Apprentissage multitâche pour la prévision de la consommation électrique

Benjamin Dubois

École Doctorale Paris-Est
Mathématiques & Sciences et Technologies de l'Information et de la Communication

Thèse de doctorat de l'Université Paris-Est

Domaine : Traitement du Signal et des Images

Présentée par

Benjamin Dubois

pour obtenir le grade de

Docteur de l'Université Paris-Est

Multi-task electricity loads forecasting

Soutenue publiquement le 2 décembre 2019 devant le jury composé de :

Jean-François Delmas	École des Ponts ParisTech	Directeur de thèse
Virginie Dordonnat	Réseau de Transport d'Électricité	Encadrante
Mathilde Mougeot	ENSIIE	Rapporteuse
Guillaume Obozinski	Swiss Data Science Center	Directeur de thèse
Jean-Michel Poggi	Université Paris-Sud	Examinateur
Peter Richtárik	KAUST	Rapporteur

Abstract in English

We study in this manuscript the **day-ahead electricity load forecasting** problem, **at the level of the substations**, based on exogenous **calendar information**, **weather forecasts** and **recent endogenous values** of the electricity demand. This work is part of a broader research field participating in the modernization of the French power system. The emergence of new production means and the evolution of electricity uses have indeed strengthened the need to anticipate the variations of the electricity demand. The **Transmission System Operator** (**TSO**), as a central actor of the electricity sector in charge of the supply-demand equilibrium and the management of the resulting energy flows, is particularly affected by these evolutions. Its decision-process relies on the ability to forecast accurately the spatial distribution of both the production and the demand. The advent of modern Machine Learning forecasting tools, in association with the improvement of computing capabilities and the gathering of rich weather and electricity datasets give rise to new opportunities.

Data exploration and the dynamic literature about electricity load forecasting serve as a basis for the extension to local forecasts of the more classical models designed for the aggregated loads. We describe a generic bivariate linear model and compare its behavior at the national and the local levels. This allows us to identify both the similarities and the heterogeneous aspects of the substations. At the local level, the data exploration and the experiments are organized around a **dichotomy** between models **learned independently** for the different substations and a **coupled modeling** of the loads. In particular, we motivate a **multi-task approach** to load forecasting with a characterization of a **common structure** encountered in the local models, that we intend to leverage for the benefit of the latter, in terms of **computational speed** and **generalization performance**.

We address several questions related to the multi-task approach. Namely, what to expect from a coupling of the local models ? Which parts of the model should be coupled and how ? How to assess the evolution and the relevance of the multi-task framework ?

We study **three coupling assumptions**, based either on a **clustering** of the model parameters, an optimization problem with a **low-rank constraint** that we analyze in details, or on the **consistency** between the forecasts at different aggregation levels. Thereby, we prove empirically that the number of parameters of the independent local models is **unnecessarily large** and we confirm the **interest of sharing** the parameters and the losses during the learning process.

Résumé en français

Nous étudions la **prévision du jour pour le lendemain de la consommation électrique agrégée à la maille des points de livraison**, à partir des **informations calendaires**, des **prévisions météorologiques** et des **valeurs récentes** de ces séries temporelles. Ce travail s'inscrit dans un domaine de recherche plus large, qui participe à la modernisation du système électrique français. Avec la pénétration des énergies renouvelables et l'apparition de nouveaux modes de consommation, le besoin d'améliorer la qualité des prévisions au niveau local devient de plus en plus pressant. Acteur central du système électrique, le **Gestionnaire du Réseau de Transport** (**GRT**) est responsable de l'équilibre offre-demande et assure en permanence la fluidité de la circulation d'électricité sur le réseau haute tension. La prévision de la consommation est donc pour lui une problématique de recherche centrale. Le développement d'algorithmes modernes de *Machine Learning*, l'augmentation des capacités de calcul et la disponibilité de grandes bases de données électriques et météorologiques laissent entrevoir de nouvelles opportunités.

Après une analyse exploratoire de la base de données et une étude de la littérature portant sur la prévision de la consommation électrique, nous considérons la possibilité d'étendre et d'adapter les modèles utilisés pour la prévision à des mailles plus larges comme les régions ou le pays. Cela nous permet de souligner les comportements plus hétérogènes au niveau des points de livraison de ces séries temporelles. L'ensemble des expériences est organisé autour d'une **dichotomie** entre des modèles à différents noeuds du réseau appris de façon **indépendante** ou bien **couplée**. Plus précisément, nous justifions une **approche multi-tâches** de ces prévisions avec les similarités entre les courbes aux différents noeuds du réseau, qui vise à améliorer la **vitesse d'apprentissage** de ces modèles ainsi que leur **capacité à généraliser**.

Nous séparons l'approche multi-tâches en trois questions. Quelles composantes des modèles est-il pertinent de coupler ? Quelles améliorations ce couplage peut-il apporter ? Comment évaluer la pertinence de l'approche multi-tâches dans le cadre de la prévision de la consommation électrique ?

Nous envisageons **trois couplages** possibles des modèles de prévision, fondés respectivement sur un ***clustering*** des coefficients des modèles, une hypothèse de **rang faible** sur la matrice de coefficients, et une mesure de la **cohérence** des prévisions à différents niveaux d'agrégation. Empiriquement, nous montrons le **caractère contingent du grand nombre** de coefficients des modèles appris indépendamment et confirmons **l'intérêt de coupler** les fonctions objectifs à minimiser ainsi que les paramètres des modèles au cours de leur apprentissage.

Acknowledgements

I sincerely enjoyed the last 3 years. Having benefited from the constant support and help of my supervisors, I found the whole project fascinating. I am therefore deeply grateful to Guillaume, Jean-François, Vincent and Virginie for their constant implication and their patience throughout this project.

I feel fortunate of having integrated the IMAGINE team and having met in this lab numerous thoughtful and passionate researchers, as well as motivated and motivating students. The position at École Nationale des Ponts et Chaussées also allowed me to discuss with people working in the CERMICS laboratory, and more generally to benefit from the vast amount of knowledge and curiosity stored in the Coriolis building, which undeniably offers the opportunity to benefit from the synergies between different research fields including, but not limited to, computer sciences and applied mathematics. I definitely feel the need to thank Brigitte and Isabelle who provided an essential support to these two research teams.

I had the pleasure of working with the Data Science Team at RTE and exchanged in various contexts with Clément, Jean, Lucie and Valentin. More generally, I am thankful to the members of the Direction Innovation et Données at RTE for their appreciated welcome.

I am also grateful for the opportunity I had of visiting the Swiss Data Science Center in 2019. I enjoyed discovering a different research environment, various applications of Data Science and am thankful to all its members for their help, their curiosity and the time I spent with them.

Last but not least, I benefited from the dynamism of the Machine Learning community, in Paris and abroad, through various events and am grateful to the organizers and the participants who make this research community live. I also admire the silent work of the open-source community that is essential to most research projects, including undeniably this one.

Thank you Guillaume, Jean-François, Vincent and Virginie, it really was a pleasurable and enriching experience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

The larger variability of the production and the demand with the general modernization of the electric power system require from the operators a more accurate anticipation and a faster reactivity. The adoption of generic statistical tools to forecast the aggregated demand and the more recent availability of data with a higher spatial resolution naturally leads to the ambition of developing electricity load forecasting models for the local demands. This work contributes to this field of research and is the fruit of a collaboration with Réseau de Transport d'Électricité (**RTE**), the French Transmission System Operator (**TSO**) in charge of the national supply and demand equilibrium and of the management of electricity flows on the high-voltage network. It is more precisely dedicated to the study of forecasting models at the level of the electrical substations.

1.1.1 The French energy system

Electricity represents 25% of the final energy consumption in France [IEA, 2016, 2019]. Per capita, the annual electricity demand is about 6.7 MegaWatt hours (**MWh**), which leads with 67 millions citizens, to a national demand between 450 and 520 TeraWatt hours (**TWh**) since 2001 [RTE, 2018]. French exportations of electricity rank globally second, after the US, thanks to the annual production of approximately 550 TWh and the country could rely entirely on self-produced electricity in terms of total consumption of energy.

History In order to structure the electricity sector and ensure self-sufficiency in terms of electricity, Électricité de France (**EDF**) was founded as a state-owned monopoly in 1946. This electricity production activity was reopened to competition in 1999 and, in 2005, company shares were proposed to the Stock Exchange in Paris. This change of status followed the European Directive 96/92/EC of December 1996, stipulating that electricity production and sales should no longer be state-regulated activities, for the benefit of the final consumer.

Structural transformation The demands in the European directive were met in the law 2000-108, to ensure the opening to competition of both electricity production

and sales, while transmission and distribution remained natural monopoly due to high infrastructural costs. They were subsequently reaffirmed in 2010 by the text *Loi de Nouvelle Organisation des Marchés de l'Électricité (Loi **NOME**)*. This led to the creation of a National Regulation Authority (**NRA**), the *Commission de Régulation de l'Énergie*, and the split of the historical electricity operator EDF, to separate the production and sales activity reamining in the EDF company from the transmission and the distribution of electricity, managed by the newly created companies RTE and Enedis (ex-Électricité Réseau Distribution France).

The EDF company started the competitive energy **production market** that it now largely dominates along with ENGIE (ex-GDF Suez) and E.ON, accumulating altogether 95% of the shares. The **transmission** of electricity on long distances is ensured by *Réseau de Transport d'Électricité* (**RTE**), the French Transmission System Operator (**TSO**), owner of the high-voltage network. After the electricity voltage is decreased below 63 kV at the so-called substations, the Distribution System Operators (**DSO**) intervene. They are engaged with local authorities through concession, maintenance and exploitation contracts for the medium and low voltage networks to route electricity from the high voltage network to consumption sites. In a given geographical area, electricity **distribution** is a natural monopoly too. The company Enedis, 100% owned by EDF, is the largest DSO and distributes electricity in 95% of the country. Local companies, the *Entreprises Locales de Distribution* (**ELD**), are in charge of the other 5%. Finally, the company EDF is also involved in the competitive **sales market**, that now counts approximatively 30 members. In short, the suppliers propose to individual consumers connected to the medium and low voltage networks sales and supply contracts. This organization of the electricity system is summarized in Figure 1.1 and accompanied by a basic representation of the largest financial and energy flows.

Fast evolution of the electricity mix While hydroelectricity was largely developed in the 1950-60's until it could provide almost 20% of the French electricity production [RTE, 2018], wind and solar energies have become financially competitive only recently, thanks to the decreasing costs of the production means, making accessible the path to a low-carbon energy system, still sustained in France by the nuclear plants that satisfy 70% of the demand.

However, **renewable energies** are uncontrollable, **hard to predict** and **fatal**, in the sense that they are either consumed instantly for free or lost since the storage of electricity is not possible on a large scale so far. While renewable energies have a legal priority on the network, the other sources of energy, namely nuclear or fossil, must adapt. The integration of the renewable energy production in the French electricity system required a modernization of the network that led, among others, to the denomination **smartgrids**. According to the ambitions set during the *Grenelle de l'environnement*, legally adopted in 2015 in the *Loi relative à la Transition Énergétique pour la Croissance Verte* and supported by various financial mechanisms, the part of the renewable energies should reach 40% by 2030.

This major evolution of the electricity mix destabilizes the supply and demand equilibrium due to the randomness of the renewable production. To ensure the safety of the energy system and attain the announced economic and ecological objectives, the anticipation of both factors, that is to say the ability to predict and meet the

demand by estimating the fatal renewable production and set the thermic (fossil and nuclear) and hydraulic production accordingly, is a key intermediary objective.

Decentralization of production Meanwhile, **a lot of local producers** of varying sizes, including many professionals producing from renewable sources, appeared with the opening to competition of the energy-related activities. Their integration in the electricity system and the measurements of the resulting energy flows is helped by the development of **smart meters**. Those have indeed become indispensable to the modernization of the networks now more generally denominated smartgrids.

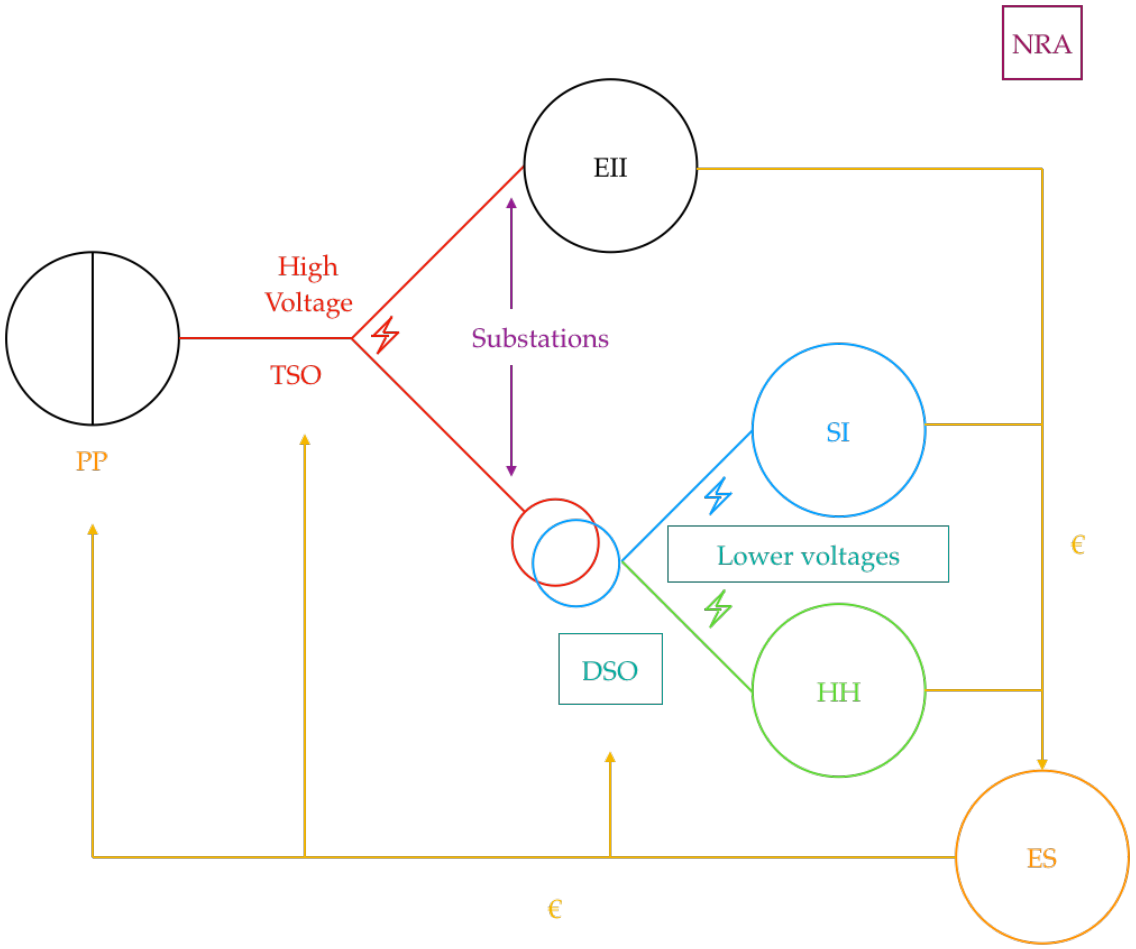
The smart meters are also responsible for an improved communication with the end-consumers, now able to better understand and control their consumption, participate in demand-response mechanisms, possibly measure local renewable production and manage the related financial flows. As a consequence, the outdated representation of the electricity system as a set of sources and sinks has been replaced by the modern networks that are the scene of bidirectional flows, both physical, financial and informational.

The additional development of non-professional local means of production modifies naturally the local electricity demand. Thereby, the multiplication of impacting decision makers in the energy system impacts the roles of the major electricity companies that used to have a stranglehold on the production plans. The **research and development** of modern tools to **anticipate and react** to the **variations of the demand** has consequently become of major interest.

The larger European ecosystem Simultaneously to the modernization of its economy, the energy system has been identified by the European Union and its members as a key component of the necessary Energy Transition in the developed countries and as a major factor of geopolitical stability.

The resulting will to mutualize the means of production and the peaks of the demand with international partners reinforces the need to consider the French energy system as a component of the larger European network. With the development of **interconnectors**, it is effectively no longer possible to isolate the country from the rest of Western Europe and the management of the network requires to also take into account the supply-demand equilibria in other countries, as illustrated in Figure 1.2. The development of these interconnectors allows to **pool the means of production** and therefore the risks, it also aims at **coupling prices** by considering Europe as a giant copper plate to make the **market mechanisms more efficient**. Among others, the satisfaction of the equilibria in the different regions requires that transboundary electricity lines are not saturated, which also demands an accurate forecasts of the needs in each region.

As exposed, the need for accurate load forecasts has become of great importance with the modernization of the energy system, in particular for the TSO, in charge of the safety and the stability of the high-voltage network.



PP	:	Power Plants	(hundreds)
TSO	:	Transmission System Operator	(only 1)
DSO	:	Distribution System Operators	(at most 1 per geographical area)
EII	:	Energy Intensive Industries	(about 260)
SI	:	Smaller Industries	
HH	:	Households	36 millions
ES	:	Energy Suppliers	(about 30)
NRA	:	Regulation Authority	(1)

FIGURE 1.1: Cartography of the French electric power system
This scheme aims at identifying the groups of actors and has been simplified as for instance, we did no plot local energy production or taxes. The 260 Energy Intensive Industries are directly connected to approximatively 2000 substations. A more detailed explanation of the energy flows is provided by [CRE, 2019] and a definition of the substation in terms of voltage is represented in Figure 1.4.

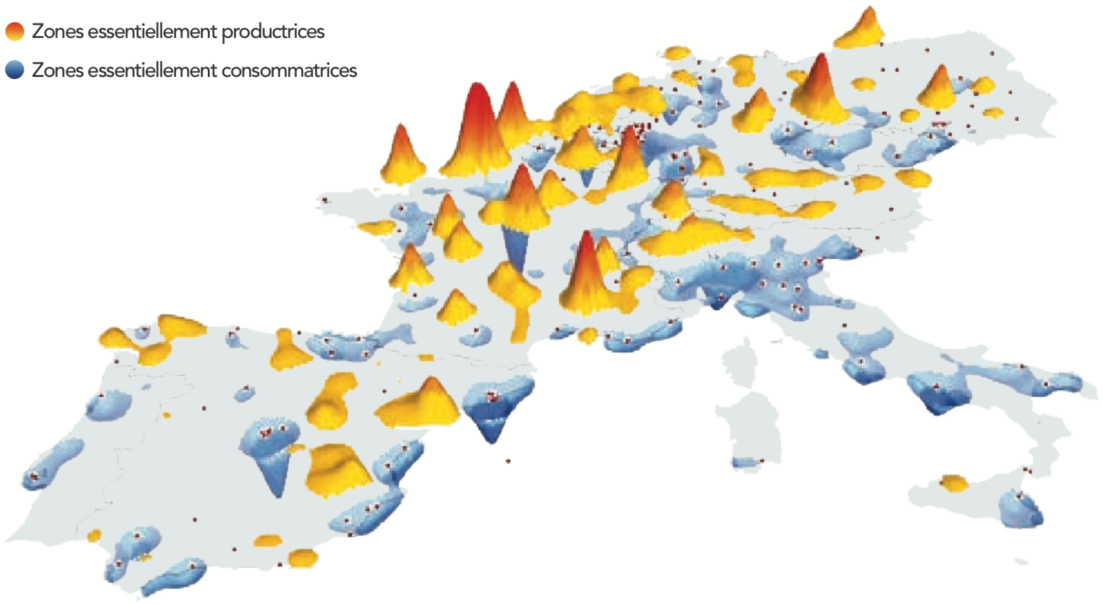


FIGURE 1.2: Spatial repartition of consumption and production
Map of Western Europe with electricity production (orange) and consumption sites (blue) [RTE, 2016b].

1.1.2 The French Transmission System Operator

Fruit of the split of the historical operator EDF, RTE is an independent limited liability corporation, still owned by public entities, notably with 50.1% of the shares owned by EDF. As a **natural monopoly**, the annual transmission of 550 TWh is a closely regulated activity, it employs approximatively 9000 people and generates an annual revenue of 4 billion euros.

Missions As a regulated monopoly, RTE is officially engaged with the French State to provide a **public service**. In particular, RTE is in charge of :

- the **supply-demand equilibrium**, which requires to inform the producers of the coming electricity demands,
- the **quality of electricity** measured by the stability of the voltage, its frequency, and the quantities of blackouts,
- the current **maintenance** and the future **development** of the high-voltage network,
- providing **support for public decisions**, in particular for the electricity tariffs and the investment programs,
- supporting the electricity markets as a **transparent, non-discriminative and independent** TSO.

Organization With a cost of about 15 cents per KWh for the end-users, the electricity annually in transit on the French high-voltage network represents about 80 billion euros. From this amount, about a third is dedicated to the regulated Tariff for Public Electricity Network Use (**TURPE**, Tarif d'Utilisation des Réseaux Publics d'Électricité), 25 % of which, that is 7 % of the price of electricity, are dedicated to RTE for the operation and the development of the high-voltage network.

1.1.3 Architecture of the high-voltage network

The high-voltage network is the physical connection between large corporates-owned production units and 4000 substations altogether.

The raison d'être Unique link between the major production sites, the local areas of consumption and the interconnectors between countries, 105 000 kilometers of electricity lines are managed by RTE to route electricity with voltages between 63 kV and 400 kV, as illustrated in Figure 1.3. As a comparison, Enedis operates 1.3 million kilometers of lines with voltages under 63 kV.

Due to the installation of local professional producers directly connected to the medium-voltage network, the former dichotomy to represent the high voltage-network with sources and sinks is outdated. However, high-voltage electricity lines are still decidedly required for the transit of electricity on long distances with minimal losses of energy.

Topology There are 2 virtual channels for the produced electricity to arrive on the RTE network from the sources. First, 54 large-scale production companies directly interact with RTE to organize their production plans. Secondly, 150 aggregators act as intermediaries between RTE and local producers.

At the other end, 4000 substations are the interface between the high and the lower-voltage subnetworks [RTE, 2019e]. Half of them supply 260 energy-intensive client companies, including 15 railway companies, notably the French National Railway Company (Société Nationale des Chemins de Fer Français, **SNCF**). The other half connects 32 DSO, managing the distribution in delimited geographical areas and routing electricity meant for residential neighborhoods, tertiary activities and small or medium-size industries.

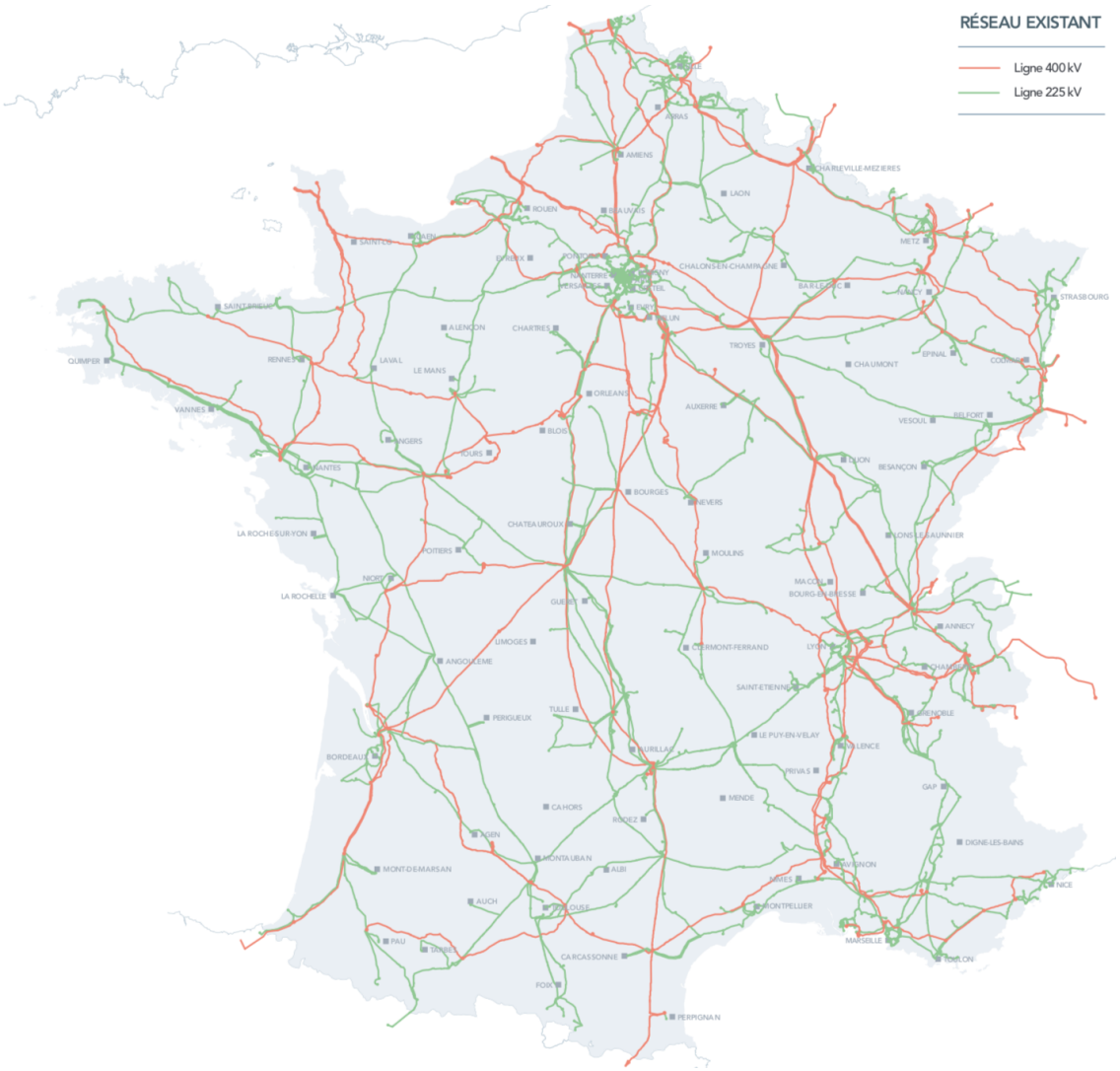


FIGURE 1.3: The French high-voltage network [RTE, 2019b].

The substations In this manuscript, our final objective is the load forecasting at the level of substations. These are defined as the approximatively 4000 interfaces between the French high voltage network and the lower voltage networks, as

illustrated in Figure 1.4. Electricity flows continuously through the substations, its intensity being directly dictated by the local demand.

The aggregated demand at the substations connecting the DSO with the high-voltage network presents relatively regular yearly, weekly and daily cycles. Besides, its strong dependence on the weather conditions and the national economic activity makes it suitable for statistical forecasting methods. On the contrary, because the demand of energy-intensive companies depends on factors significantly different and hard to model on a large scale, they are not considered in this manuscript.

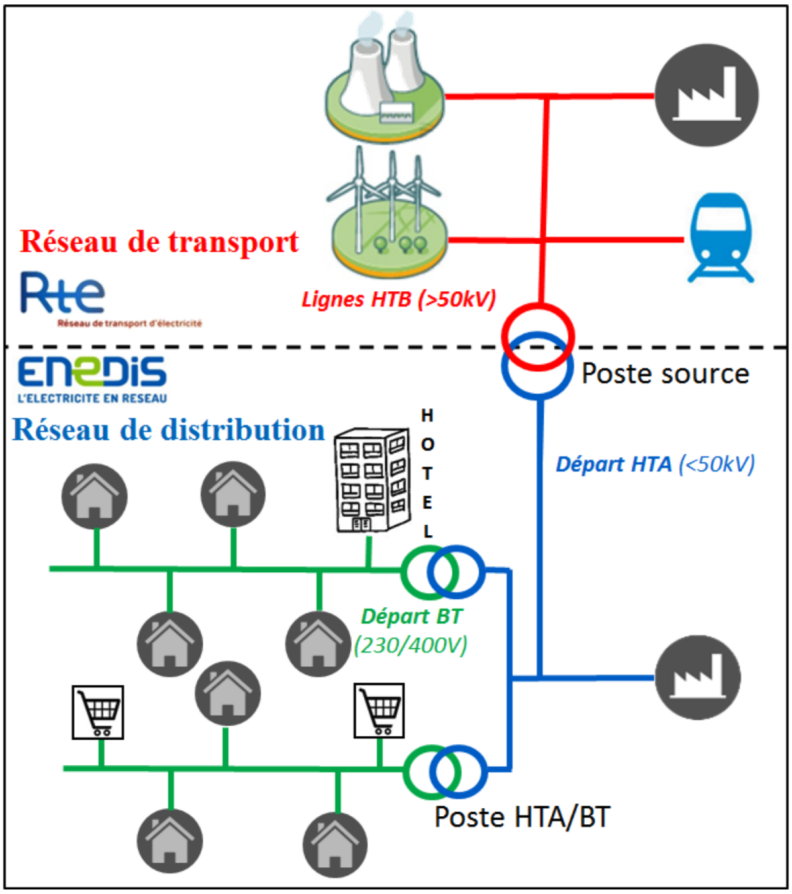


FIGURE 1.4: High and lower voltages networks
Illustration of the substations (*Poste source*) as the interfaces between the high-voltage network and lower voltage areas [Barbier, 2017].

1.2 Load Forecasting

Like the aggregated energy demand, the electricity consumption is sensitive to the economic context. In 2016, the French energy intensity was estimated around $0,12$ ton of oil equivalent per 1000 euros ($1,4$ MWh per 1000 euros) and the GDP was 2225 billions euros.

It is also sensitive to the weather conditions : the slope of the electrical load with respect to the temperature is around $\frac{d\ell}{dT} = -2.4$ GW/°C in winter [RTE, 2016a] because of the increased heating demand, which corresponds to a relative variation of $\frac{100}{\ell}\frac{d\ell}{dT} = -2.7$ % /°C, and $+0.4$ GW/°C in summer mainly due to the presence of cooling appliances.

Estimating the coming electricity demand and adapting flow-management accordingly is a key step for RTE to carry out the missions entrusted by the State. There are for RTE multiple load forecasting problems to consider, each being characterized by : an aggregation level, between the whole country and the substations, a time horizon ranging from few minutes ahead to several years in the future, and a temporal granularity, for instance every couple minutes for intra-day forecasts and every couple hours for the loads in a decade.

1.2.1 Day-ahead local load forecasting

The day-ahead national load forecasting problem has been studied by the research community for several decades and a forecasting model has been operational at EDF since the 1980s. In this manuscript, we focus on the day-ahead load forecasting problems, at the **local** level of substations.

More precisely, we consider the problem of **day-ahead deterministic hourly forecasts of the local load** at every substation, meaning the forecast at 23:59 on day j of the hourly loads on day $j+1$, with a preference for an interpretable model given that eventually its usage should not be restricted to statisticians. With the 2 000 substations considered, this corresponds to 48 000 values to forecast everyday.

1.2.2 Industrial interest

Facing the increased variability of the supply and the demand, a predictive tool is a prerequisite for the local management of power systems to ensure its stability and its resilience. The penetration of electric vehicles and the installation of renewable energy power plants are only 2 of the major challenges to come in the next decade. Altogether, we identify 4 main needs for the TSO to set a local load forecasting model.

National supply and demand equilibrium RTE is contractually responsible for the national supply and demand equilibrium. This is part of its public service missions agreed with the French State and monitored by the National Regulation Authority (**NRA**) of the energy sector, that justify the financial compensation known as the TURPE. To this end, RTE has a clear interest in anticipating the load and the fatal renewable energy production *i.e.* the wind and solar productions. Otherwise, RTE is assisted by the so-called *responsables d'équilibre* (responsible for the equilibrium), that agreed to finance, against a predetermined remuneration, the difference

between the electricity injected on the high-voltage network and the electricity effectively consumed.

Safety of the system flow management The equilibrium must also be satisfied locally to ensure the feasibility, determined by the capacity of the lines, of the production planning and the electricity transfers. The resilience of the network is commonly assessed by its capacity to resist the default of a couple random electrical lines. This requires in particular to estimate the future loads at different crucial nodes of the system including the substations, key interfaces between the high-voltage network and medium or low-voltage networks.

Maintenance planning Maintenance is a necessity for RTE to ensure the safety of the network. Either predictive or corrective, it often requires to disconnect electrical lines. In such a case, the load forecasts permit to check the feasibility of the energy flow planning and ensure the robustness of the network, in spite of the supposedly offline part of the grid.

Loss reduction Finally, the anticipation of electricity demand is necessary to consider the optimization of energy flows and losses, in terms of distance travelled between the production and the consumption sites illustrated in Figure 1.2. The losses due to the Joule effect are indeed more or less proportional to the distance traveled. Although they are reduced thanks to the high voltage, they oscillate between 0.7 GW and 3 GW. Note that this physical phenomenon is not the only cause for energy losses on the network, the iron losses occurring in transformers being of the order of 0.1 GW.

1.2.3 Original motivation of this work

In addition to the importance of accurate forecasts of the local loads in the decision-making processes of RTE, this work is generally motivated by the current dynamic of the Machine Learning community working on forecasting models [Hahn et al., 2009; Kyriakides and Polycarpou, 2007; Muñoz et al., 2010; Weron, 2007], the new availability of large datasets [Hong and Fan, 2016; Hong et al., 2014] and the accessibility to more computational power, making precisely possible the consideration of these datasets. The important impact of the temperatures on the electricity demand also makes essential the quality of local weather forecasts, constantly improved during the last decades.

All the factors above motivate the common idea that the results on existing forecasting problems can be improved thanks to the development of modern scientific tools. However, we explain in this manuscript that the local load curves can be significantly different from the national or the regional loads since they do not benefit from the same smoothing effect due to the aggregation. Therefore, their volatility is higher, even though a lot of similarities can be observed between the substations. They have also been less studied and their relationships with the weather and the calendar information are not as well understood. As a first consequence, the models developed to predict the national load may be inadequate at the local levels. Sec-

ondly, computational power is not the only extra ingredient required for local load forecasting.

Instead, the ambition of this work is to characterize the similarities observed at the level of substations and propose a modeling able to benefit from them, in terms of numerical accuracy and computational time. This manuscript addresses the following question :

Is it possible to leverage the similarities between local loads to improve the forecasts with coupled models ?

Organization of the manuscript In Chapter 2, we present the database provided by RTE and Météo-France in order to specify the problems we are interested in. We also present related work and a preliminary data exploration allows us to justify our approach.

Chapter 3 is dedicated to load forecasting models where each time series is dealt with independently from the others. We propose a modeling based on B-splines for univariate effects and products of B-splines for bivariate effects, which leads to a standard bivariate linear model. After casting and solving the optimization problem both at the national and at the local levels, we propose an analysis of the results to highlight the difficulties encountered in the modeling and justify the multi-task approach of Chapter 4.

Illustrating the models learned and their residuals in the independent setting lets us relate the local load forecasting problem with different multi-task approaches presented with related work at the beginning of Chapter 4. We study three different multi-task models. The first one assumes that the coefficient of the models for different substations are close in a geometrical sense. It is based on a clustering method. The second approach is geometrical too but only assumes that the coefficients learned for the different models lie in a low-dimensional space. This leads us to considering an optimization problem with a low-rank constraint, which is studied in details in Chapter 5. The analysis of the convergence to critical points and the linear convergence in a neighborhood of the optimal set has been published at the International Conference on Artificial Intelligence and Statistics in 2019. Finally, the third approach presented in Chapter 4 has two motivations. It is both an attempt to leverage the correlations between the residuals observed with the independent models and a proposition to have local forecasts consistent with the aggregated forecasts, at the national or regional level, meaning that the sum of the local forecasts must be a reasonable forecast of the aggregated electricity demand.

Chapter 2

Data exploration and methodology

The database contains load measurements at 2089 electrical substations and 32 weather stations distributed all over the 12 metropolitan regions of France (Corsica is not part of RTE network) and presented in Figure 2.1.

In this chapter, we present the target variables along with their relationships with the classical inputs of load forecasting models. We also present the ambitions of this work and the methodology that we adopted.

The choices of the notations for the manuscript are explained in Appendix A.

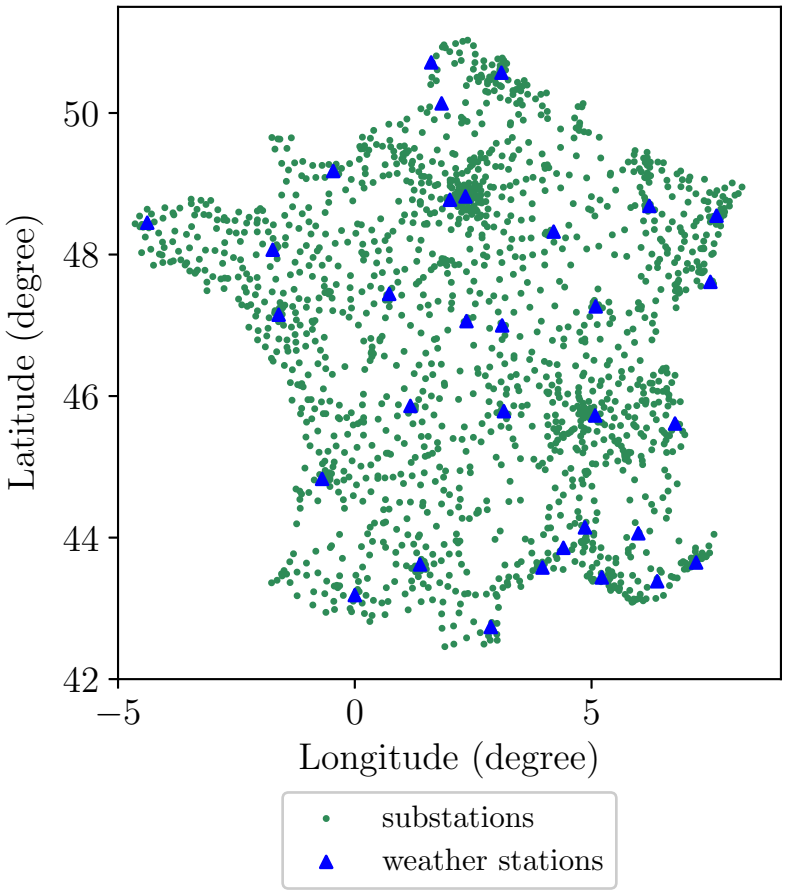


FIGURE 2.1: Map of all substations and weather stations.

2.1 Raw load data

The loads are measured by the TSO at the level of substations, defined with Figure 1.4, and are given in MegaWatt hour (**MWh**) with UTC time.

Discarded special clients Originally, there are about 4000 electrical substations in France. Half of them, connected to the so-called *special clients*, only serve energy-intensive consumers like the French National Railway Company (**SNCF**), or specific industrials producing mainly steel, aluminum, glass, paper, cement or chemicals. The other half serve residential areas with possibly smaller industries and tertiary activities. The load curves of the *special clients* are very different from the other half of the substations, they exhibit specific behaviors that make their demand non-suitable for prediction, or at least requires specific attention. As a consequence, we are only interested in this manuscript in forecasting the loads of the second group of substations that are more regular and homogeneous. Discarding the *special clients* brings down the number of substations to 2089.

We should also emphasize that for this reason, the sum of the loads at these 2000 substations does not equal the national load that is made public by RTE on the website Eco2Mix [RTE, 2019a]. While the mean of the former is approximately 40 GWh, the mean of the latter is about 60 GWh. In particular, the Eco2Mix load corresponds to a larger aggregation and most forecasting models will obtain better relative performances with this time series, even though it includes *special clients* whose demands are difficult to forecast individually.

Data collection procedure At the level of substations, what is measured is the amount of electricity transiting from the high-voltage network to the lower-voltage networks. It does not take into account the locally produced electricity whose importance is growing, that is injected in the middle-voltage network and that can be considered as never transiting on the high-voltage part of the network. Yet, the TSO is interested in estimating the local demands and therefore set up a procedure that we present below, to take into account these local productions. Note that we also added a correction procedure described in Section B.1, to detect and possibly correct anomalous values, leading to the supplementary elimination of 300 substations : there are altogether 2 correction steps and the final number of substations considered is 1751.

Consider a given substation serving a delimited area. We denote $P \in \mathbb{R}_+$ the local production and always assume that it is consumed in that area. We denote by $\ell \in \mathbb{R}_+$ the total quantity of electricity effectively consumed in that area and by $T \in \mathbb{R}_+$ the electricity transiting through the substation from the high voltage network to the lower voltage network. We have the relationship :

$$\ell = P + T. \tag{2.1}$$

The quantity of interest throughout this manuscript is ℓ but what the TSO measures at the level of substations is the quantity T. Therefore, the TSO set up a procedure to estimate the unknown quantity P.[1]

[1]The local production P may be known by the Distribution System Operators but this information is not shared.

The local production P corresponds mainly to wind and solar generated electricity with relatively small installations. To estimate P in a given area, the TSO measures the quantity of renewable energy produced in nearby large installations that are effectively connected to the high-voltage network and extrapolates based on expert knowledge.

Although this estimation is imperfect as explained in Appendix B.1, it leads us to consider the supposedly simpler forecasts of the total local electricity demand ℓ instead of T which is much more sensible to the solar radiation, the wind and the local means of production. Forecasting T is a possible extension of this manuscript.

Size of the database. The collected database contains hourly measurements from January 2013 to December 2017, that is 5 years of data. It corresponds to 43924 observation instants and saved as a *csv* file, it occupies 1.8 GigaBytes. With modern computers, storing this database in Random Access Memory (**RAM**) is never a limiting factor.

Remark 1. *The size of the file containing the loads could be significantly reduced since they are measured with 15 decimals i.e. down to the NanoWatt hours. We do not however, consider such memory problems in this manuscript or try to fit models with lower precision data types.*

Limited history The decision to consider a 5-year-long database is not arbitrary. The cost of gathering reliable measurements is not prohibitive and it would be possible to train and test forecasting models on longer periods. However, the load time series are not stationary and this is particularly visible at the level of substations. We identified two main reasons for this non-stationarity.

First, the configuration of the network and the evolution of the local electricity demand evolves faster at the local levels than at the national level since it is more sensible to variations of the demography or of the economy with the arrival of new small industries in a given area. Secondly, the correction procedure set up by the TSO and described in Section 2.1 is rapidly evolving with the introduction of new means of production from renewable sources of energy. The database is therefore limited to the years 2013 to 2017.

2.2 Raw meteorological data

In this section, we present the weather information provided by Météo-France that is part of the inputs in load forecasting models. The electricity demand is indeed particularly sensitive to the weather conditions, heating being one of the major uses of electricity in France.

Origin of the data As the French national meteorological service, Météo-France provides forecasts and actual measurements with different spatial granularities and time horizons. Among others, temperatures (°C) and cloud cover indices (ranging from 0 to 8) were gathered in a dataset describing the weather at 32 geographical locations, which we call the RTE panel, roughly covering the French metropolitan

regions. The Voronoi diagram of these 32 weather stations is presented in Figure 2.2.

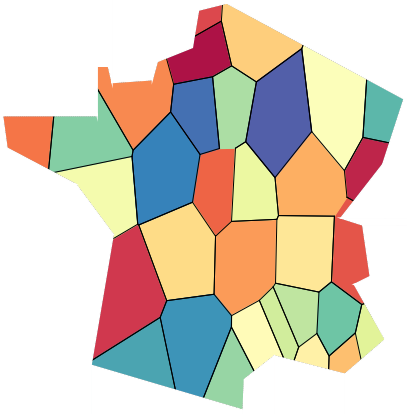


FIGURE 2.2: Voronoi diagram of the 32 weather stations

Clean dataset The weather data has been cleaned before by Météo-France. Consequently, we do not use any correction mechanism. Besides, all the experiments in this manuscript were made with observed weather conditions measured in UTC time. In operational conditions, we should of course use weather forecasts instead but we have considered that doing this to study forecasting models would increase the level of noise in the data and make their interpretation more difficult while it would not change the conclusions.

Weighted average of the weather conditions Orignally, RTE used the 32 selected weather stations to compute a weighted average injected in the national load forecasting model. The corresponding vector of weights $\boldsymbol{\alpha}_{\text{national}} \in [0,1]^{32}$ is given in Table F.1.

For each month in the dataset, the distribution of the weighted mean of the temperatures computed with the weight vector $\boldsymbol{\alpha}_{\text{national}}$, is illustrated in Figure 2.3. The distribution of the temperatures in winter and in summer is of particular interest as it determines the range of observed values. The same box plots for the cloud covers are given in Figure F.1.

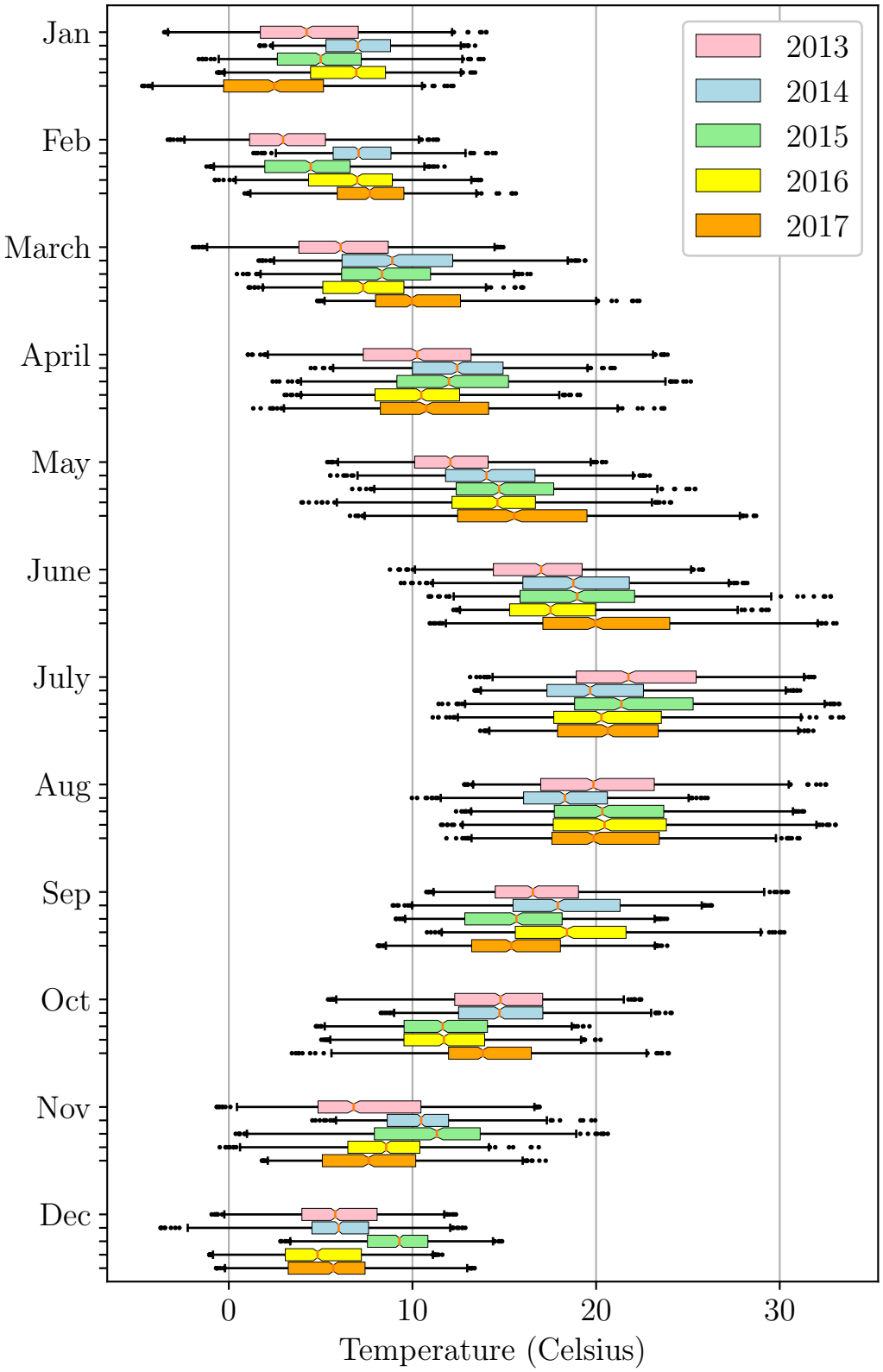


FIGURE 2.3: Distribution of the average temperatures
Box plots of the weighted average temperatures for each month in the dataset. The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extends from the 1st to the 99th percentiles. Points outside these bounds are plotted individually.

2.3 Exploratory analysis of the national load

The load can be linked with weather conditions and economic or demographic circumstances. These circumstances are indeed responsible for general trends and cycles of the loads as well as irregularities and abrupt changes. Pierrot and Goude [2011] classified these relationships into 3 groups : social, natural and economic. In this section, we review these relationships and illustrate them with the database, in order to justify our choices of inputs for the models in Chapters 3 and 4. We also highlight the non-stationarity of the load time series.

Note that the variables that we present in this chapter are not independent. In particular, the hour of the day and the position in the year are strongly correlated with the temperatures. Therefore, all the marginal distributions estimated in this section should be interpreted with caution.

2.3.1 Impacts of natural events

While a high diversity is suspected among the ways the meteorological conditions can impact the electricity demand, the pieces of information related to the weather that provably improve the quality of forecasting models are relatively restricted. They are mainly the temperatures and the cloud covers.

Effectively, the distribution of the loads illustrated in Figure 2.4 is highly correlated with the distribution of the temperatures presented in Figure 2.3. In November, December and January, the loads were larger in 2013 and 2017, which is consistent with the low temperatures plottted in Figure 2.3 and the fact that the weather is one of the main driver of the electricity demand. Note also that the distributions of the loads over different years appear much more heterogeneous in winter than in summer.

Additionally, while the wind speed at low altitudes and the humidity are sometimes included in load forecasting models, their effect are relatively small and we have chosen not to include them. This is discussed in more details in Section 3.7.3.

Instantaneous temperatures The effect of the temperature on the electricity demand is relatively well understood. Heating is roughly responsible for half of the electricity consumption in France every year.

In winter, the national load increase after a 1°C decrease has been stable over the last few years and was estimated by RTE around $2,4$ GW [RTE, 2016a, 2019c]. All things equal otherwise, the demand is minimal when the outside temperature is close to 19°C. Finally above 25°C, the electricity demand slightly increases anew, about +500 MW/°C [RTE, 2019d], due to the use of cooling appliances, thus giving the shape of a hockey cross to the marginal load curves in Figure 2.5.

Depending on the region and its usual exposure to low and high temperatures, the number of heating and cooling appliances as well as their efficiency may vary. As a result, temperature variations impact the electricity demand differently. This is illustrated in Figure 2.6.

Past temperatures The past temperatures may also impact the present loads. First, the thermal masses of the buildings make the past temperatures responsible for

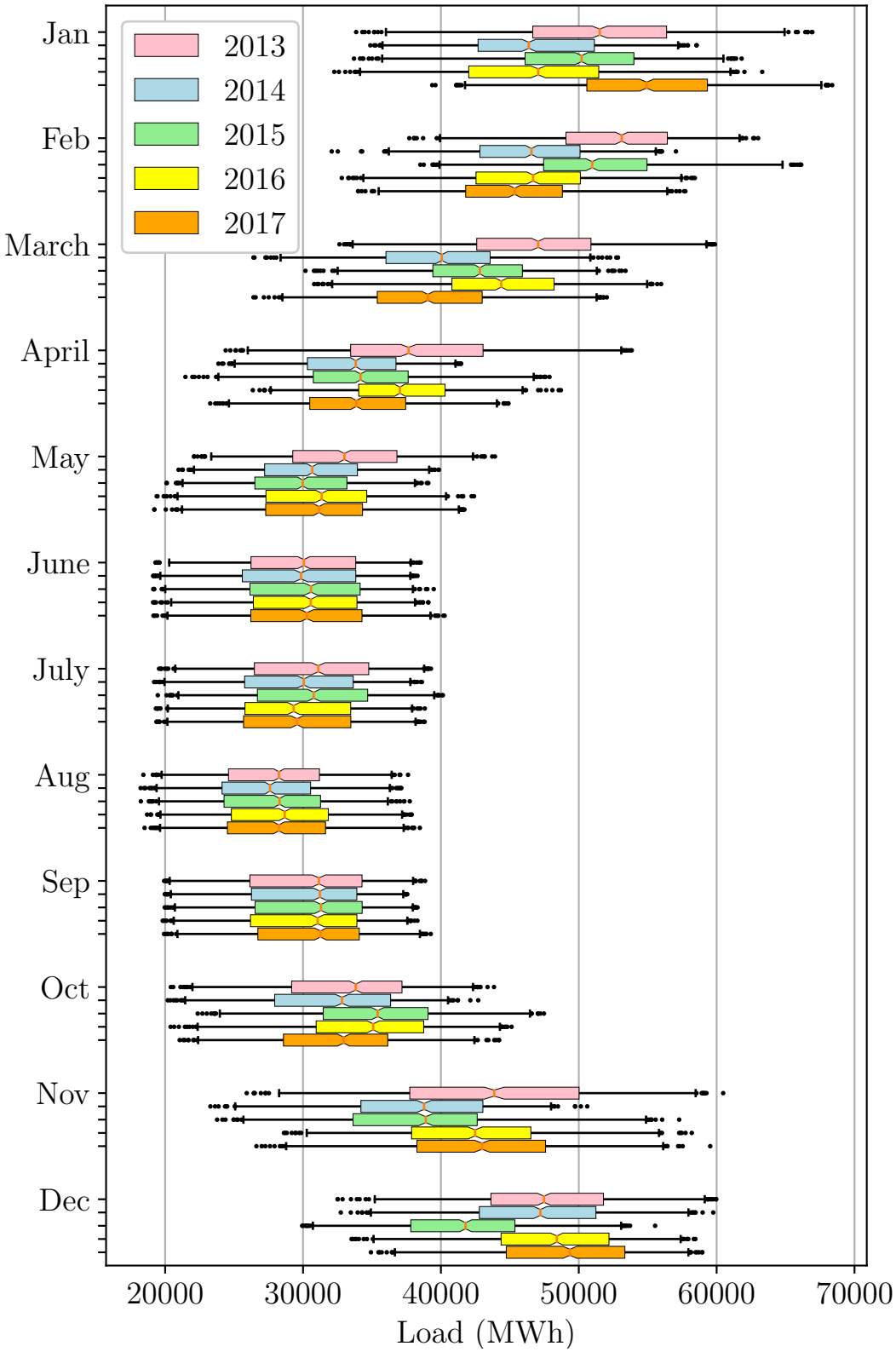


FIGURE 2.4: Distribution of the aggregated load
Box plots of the aggregated load for each month in the dataset. Note the increased variability in winter and in particular the correspondence between the largest demands and the coldest temperatures in Figure 2.3.

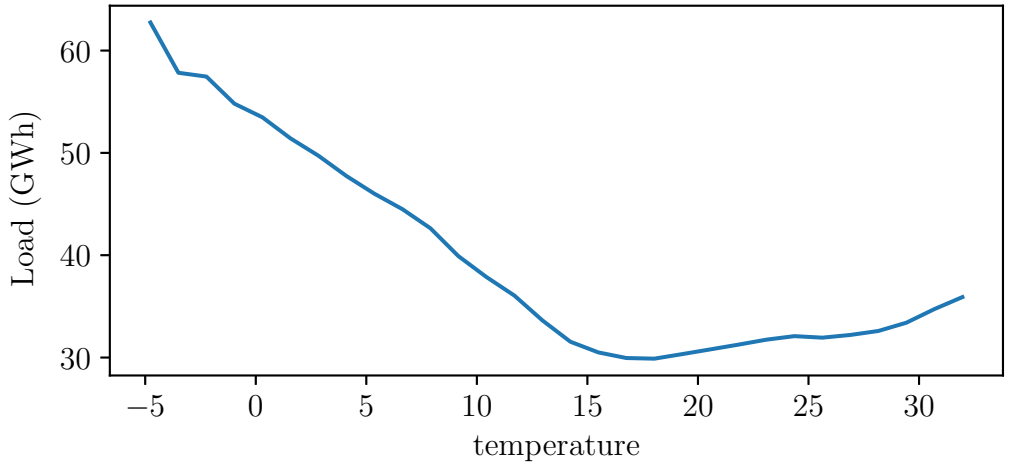


FIGURE 2.5: Average load at a given temperature value
Empirical mean of the national load conditioned on the average temperature. The corresponding density is given in Figure F.5.

the present heating demand. Secondly, people may not react instantaneously to the temperature variations. It is therefore relevant to include the past temperatures in the load forecasting models. Plus, it seems reasonable to always use the observations and never the past forecasts, even in operational conditions.

Cloud cover By impacting the need for artificial light and the heating of the buildings with solar radiations, the sunlight and indirectly the cloud cover also drive the electricity demand. To account for this effect, we have included in the modelings the cloud cover indices provided by Météo-France. We also use a nationwide indicator of the presence of the sun above the skyline in Paris but we do not have information about the intensity of the sunlight. The interactions between the cloud cover and the indicator of the daylight is of particular interest, if we consider that the cloud cover has no influence on the demand at night. The average load conditioned on the value of the cloud cover index is illustrated in Figure 2.7.

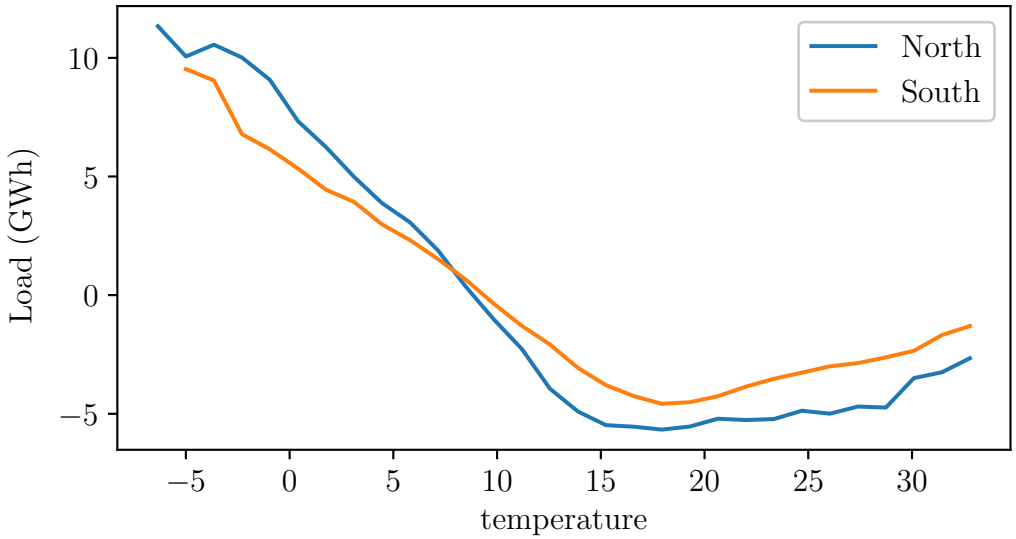


FIGURE 2.6: Load for different temperatures in the North/South
Empirical mean of the load conditioned on the average temperature in two different areas : North and South of France. The substations and the weather stations were split into two balanced groups, depending on their latitudes being above or below 46.5°N. For each group, we compute the average loads per substation per quantiles of temperature. Each of these conditional means is then centered, as a function of the temperature, to illustrate that the slope of the average load is globally larger in the South : more positive for warmest temperatures and less negative for cold temperatures. These differences may be due to the different appliances installed in each region but a definite causal interpretation is difficult.

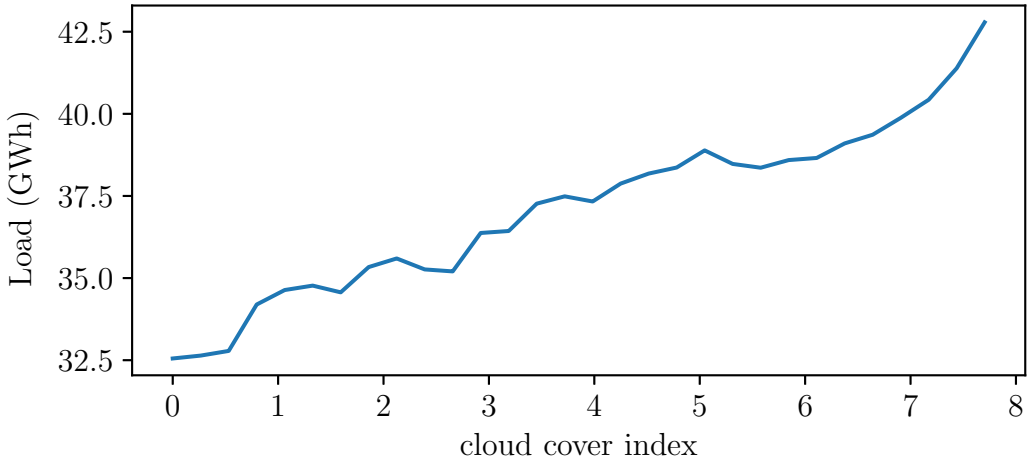


FIGURE 2.7: Load conditioned on the cloud cover index
Empirical mean of the load conditioned on the average cloud cover. The corresponding density is plotted in Figure F.6. The value 0 corresponds to an absence of clouds and the value 8 to a very cloudy observation. Note that the variations of the load should not be attributed entirely to the cloud cover since the latter is highly correlated with the temperatures, among others.

2.3.2 Impacts of the economic activity

In addition to the weather conditions, the economic activity is another important driver of the electricity demand. While it is not possible to measure people's occupations, their relationships with the calendar variables such as the hour and the positions in a week or in a year can be leveraged to estimate it.

Daily cycles The hour of the day is an obvious indicator of the level of economic activity and this is reflected immediately by the electricity demand. A similar pattern of consumption, illustrated in Figure 2.8, is roughly repeated every day.

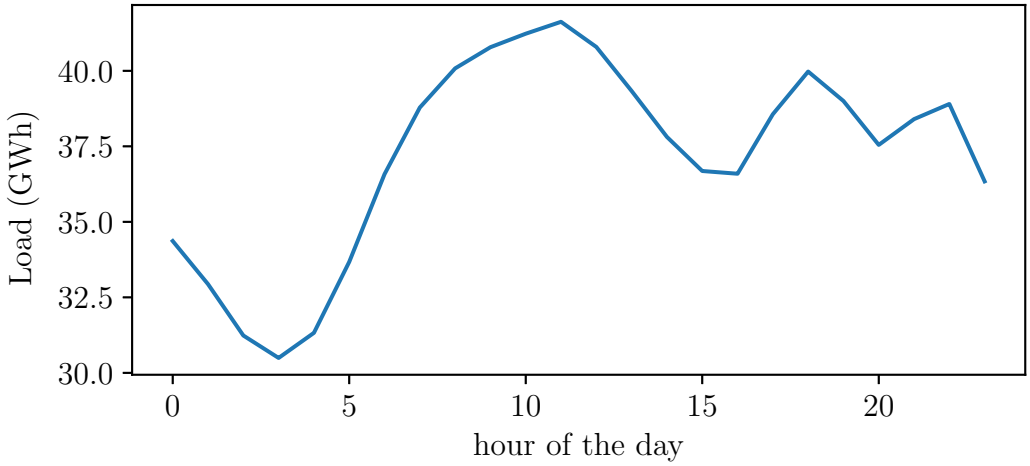


FIGURE 2.8: Daily cycles
Empirical mean of the nationally-aggregated load conditioned on the hour of the day (UTC). The corresponding density is plotted in Figure F.2. The second peak in the evening around 10 pm corresponds to the switch to the off-peak rates of electricity.

The amplitudes and the precise hour of the peaks in the morning and in the evening may depend on the period of the year, as illustrated in Figure 2.9. This is due among others, to the switch between Standard Time (**ST**) and Daylight Saving Time (**DST**) and to the variations of the sunrise and the sunset times.

Weekly cycles Similarly, the day of the week roughly determines people's agenda. Based on preliminary experiments, we believe that it is actually more relevant to consider in the modeling directly the hour of the week which is the combination of the hour of the day with the day of the week and has $24 \times 7 = 168$ possible values. The weekly cycles are illustrated in Figure 2.10.

Yearly cycles Figure 2.5 illustrates the relationship between the temperatures and the load. Since the temperatures is strongly correlated with the day of the year, we observe yearly cycles in Figure 2.11. Contrary to the temperature that is minimal during winter and maximal during summer, the electricity demand is on average maximal during January-February and minimal during July-August.

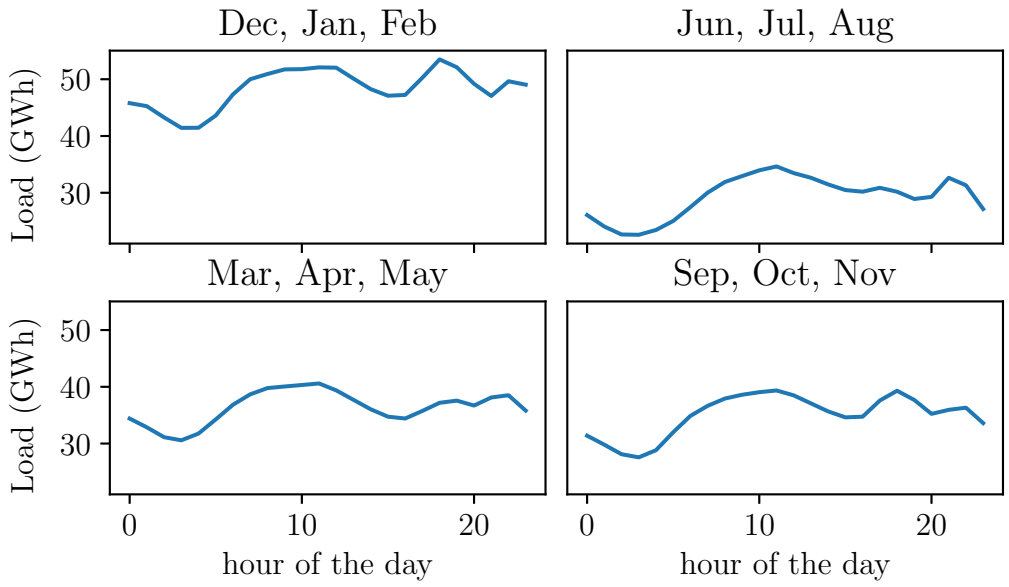


FIGURE 2.9: Daily cycles per quarter
Empirical mean of the load conditioned on the hour of the day for the 4 quarters.

It is possible that the variations of the loads along the day of the year is not only due to the variations of the temperatures. Empirically, it is useful to include both the temperatures and the day of the year in the inputs of forecasting models.

Vacations periods Of course, the variations of the temperatures are combined in summer with a slow-down of the economy which leads to the minimum loads around mid-August. A similar phenomenon occurs during the Christmas period and this is visible in Figure 2.11.

Holidays Depending on the day of the week, holidays may also affect the previous and the next days. For instance the date of Easter is not fixed and it makes its effect difficult to anticipate. Holidays can even be associated with 3 or 4-day-long weekends and affect the load in the surrounding period. We have in our database the 11 French holidays every year but no local holiday was taken into account. For instance the nationally aggregated load, at the beginning of May 2013 where 3 public holidays occur, is presented in Figure 2.12.

Long-term trend In addition to the 3 cycles identified previously, namely in a day, a week and a year, non-stationarity can also be illustrated with the observations of several years of data. This evolution results from variations of demography and economic activity at a national level, among others. It deserves a special treatment in state-of-the-art models. However, the nationally-aggregated annual demand stagnated between 2013 and 2018 and no clear evolution can be distinguished in Figure 2.13. It is more important at the substations level as explained in Section 2.4.1.

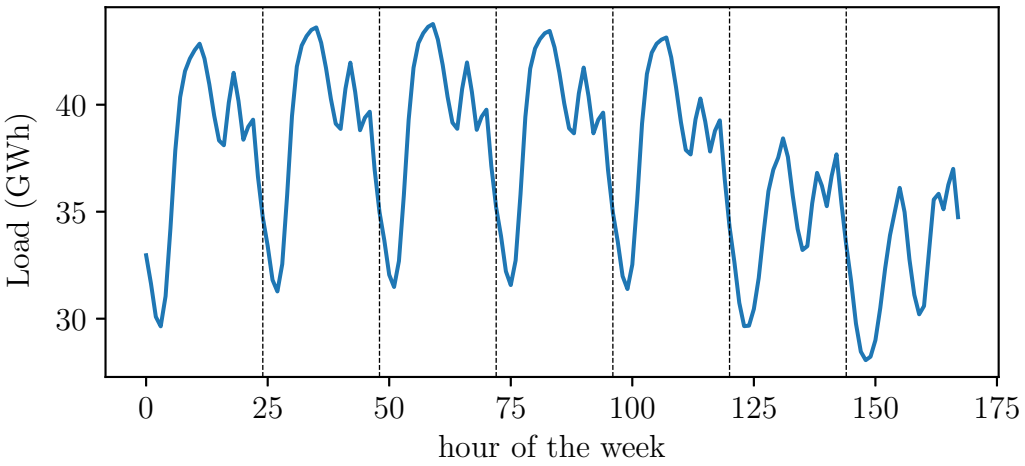


FIGURE 2.10: Weekly cycles
Empirical mean of the load conditioned on the hour of the week. The last value 167 on the x-axis corresponds to Sundays at 11 pm and the vertical lines separate the different days of the week. The corresponding density is given in Figure F.3.

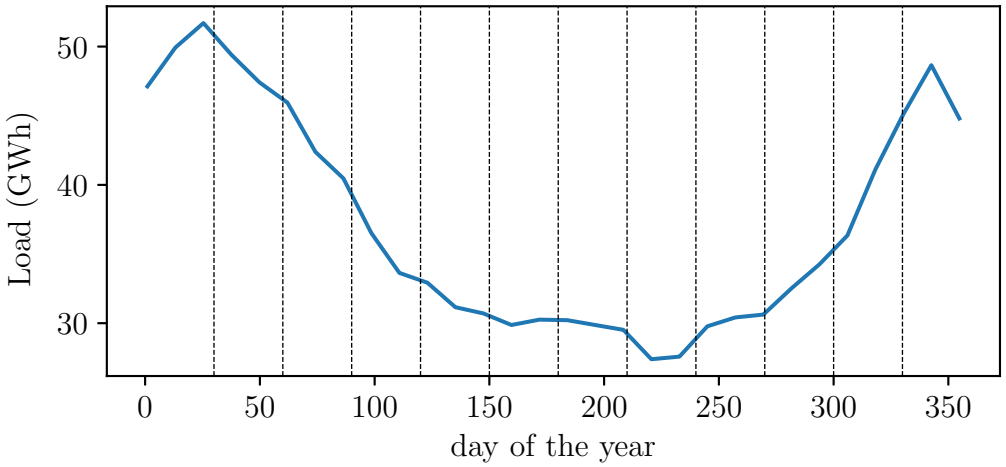


FIGURE 2.11: Annual cycles
Empirical mean of the load conditioned on the day of the year. The vertical lines separate the different months. The corresponding density is given in Figure F.4. We can observe a decrease of the activity around the Christmas period (day 350 to day 5) and during the summer break (day 200 to day 240).

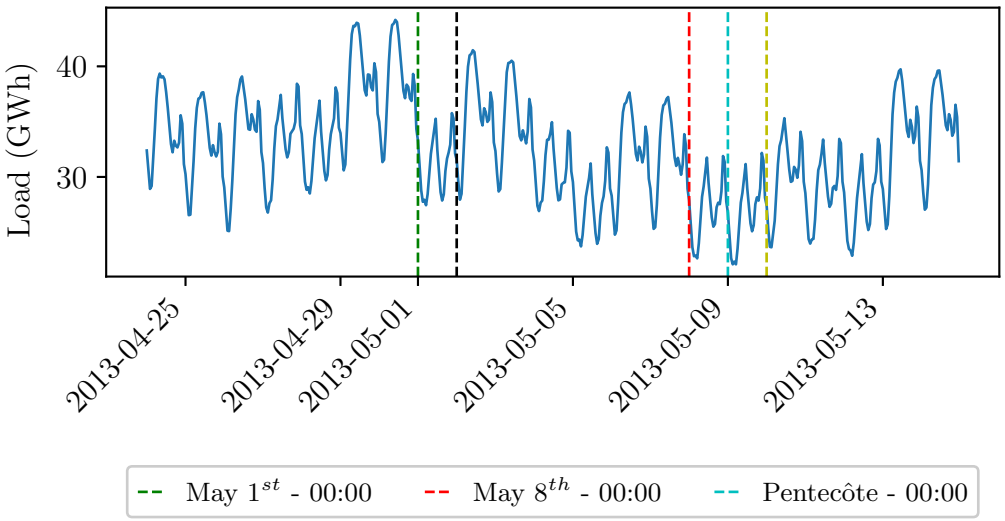


FIGURE 2.12: Public holidays in May 2013
The public holidays in May 2013 were : Wednesday, May 1^{st}, Wednesday, May 8^{th} and Thursday, May 9^{th} that was the Pentecost.

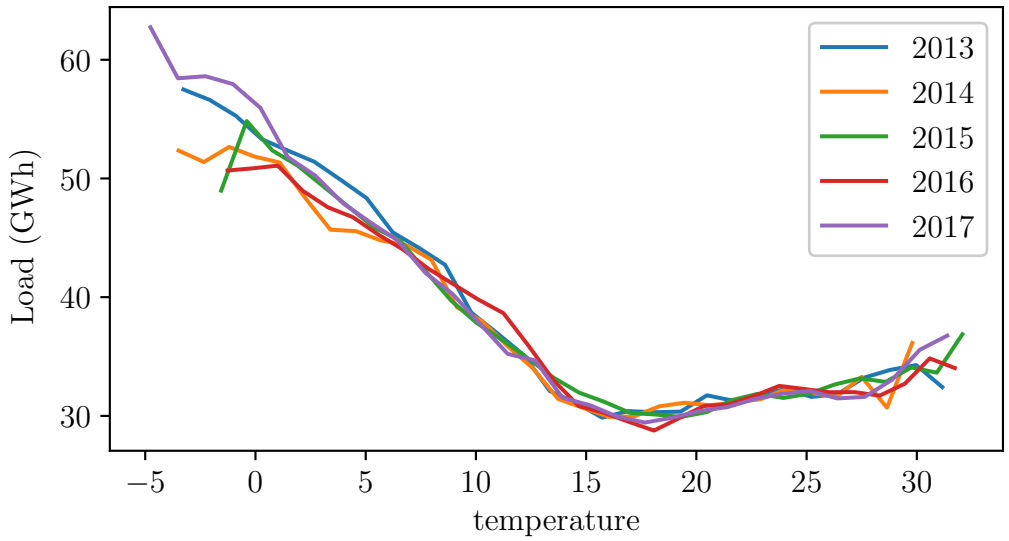


FIGURE 2.13: Long-term trend
Marginal loads per quantile of temperatures for each year in the database.

2.3.3 Notable bivariate conditional expectations

A precise modeling of the loads also requires to take into account the potential interactions between the inputs. The empirical expectations of the nationally aggregated load conditioned on the hour of the week and the day of the year is presented in Figure 2.14. Like in Figure 2.9, we observe that the peaks and the valleys during the different days of the week occur at different times over the year. It is a clue but not a proof that this interaction is relevant for the modeling of the electricity demand.

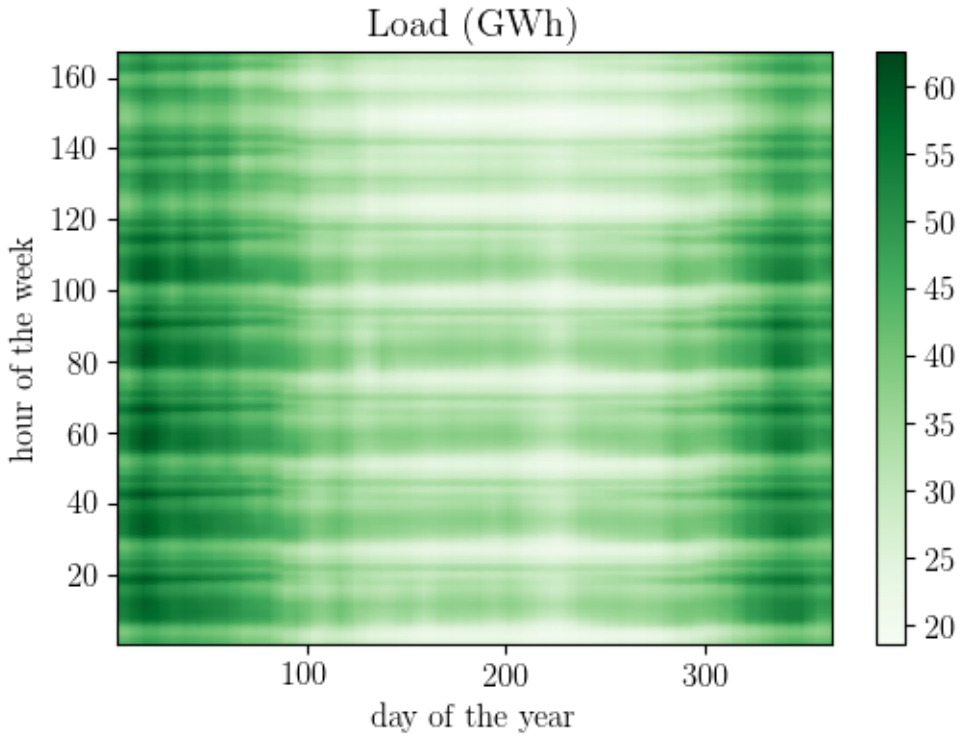


FIGURE 2.14: Load conditioned on hours and days of the year
Expectation of the nationally aggregated load conditioned on the hour of the week and the day of the year. Note that the shift during the year of the peaks in the morning and in the evening described with Figure 2.9 are also visible in this illustration. They are particularly visible but not limited to the transition to Daylight Saving Time around days 90 and 300.

The interactions between the other pairs of inputs are presented in the seven Figures F.7 - F.13.

2.4 Exploratory analysis of the local loads

Being computed as a sum of the local loads, the national load curves benefit from a smoothing effect attributed to the law of large numbers. We have highlighted in Section 2.3 the resulting regularities visible on the national load curves and the conditional expectations with respect to the different inputs. At the level of substations, most of these characteristics remain but can be greatly perturbed. In this

section, we first insist on the heterogeneity among the substations but we highlight in a second part the presence of similarities.

2.4.1 The local load curves are more erratic

We illustrate in this section the behaviors of the local load curves. We expect a higher variability at the level of the substations, compared with the national load illustrated in Section 2.3.

Higher variabilities Since aggregating the loads at the national level induces a smoothing, one must expect a higher volatility of the load curves for thinner aggregation levels. As a first illustration of this behavior, we propose a regression of the standard deviations of the local loads on their mean values in Figure 2.15. The slope 0.26 must be compared with the ratio 0.24 of the mean load that is 39 GWh at the national level with a standard deviation of 9.49 GWh.

In fact, we believe that Figure 2.15 accounts more for the amplitude of the yearly cycles than for the erratic character of the local curves. If instead of considering the standard deviations of the load curves over 5 years, we consider the average of the standard deviations within each week of the database, a similar regression gives a slope of 0.15 for the substations, 0.13 for the administrative regions and 0.12 for the nationally-aggregated load. Although, this corroborates the idea of more chaotic local loads, one could again consider that this accounts mainly for the amplitude within each week. Computing these standard deviations after subtracting an estimator of the loads that depends on the hour, the day of the week, the day of the year and possibly the temperatures would be more convincing but we defer the models and the analysis of their redisuals to Chapter 3.

Increased nonstationarity The slowly evolving trend observed at the national level is much more visible at local levels. The repartition of electricity uses evolves relatively faster in smaller areas, thereby modifying the joint distribution of the load and the inputs.

As an illustration we regress the monthly loads with the absolute time. At the national level, the computed relative slope is -1.31 % per year and the distribution of the slopes computed for the substations is illustrated in Figure 2.16. Note that this regression might be disturbed by boundary effects and in particular the fact that the dataset is only 5 years long. We discuss the non-stationarity in more details in Section 3.7.4.

Importance of holidays and vacation periods Depending on their location and the categories of the connected clients, the substations can be much more sensitive to the holidays than the national load. We present in Figure 2.17 the load curves of 2 substations, one whose behavior is relatively similar to the national load and the other that is significantly impacted by the holidays, the summer break and the Christmas period.

Empirical conditional expectations The conditional average we presented in Section 2.3 are also significantly different at the level of the substations. While the

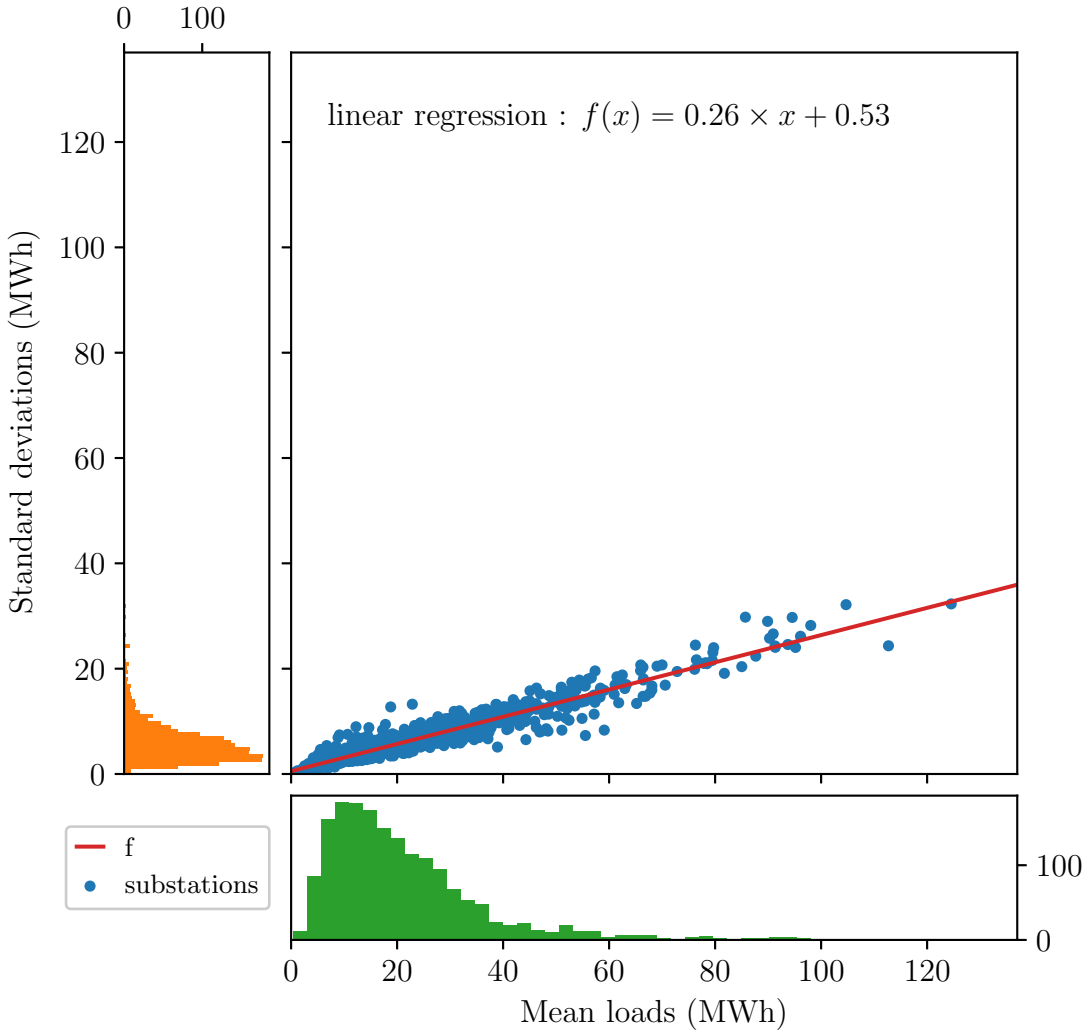


FIGURE 2.15: Means and standard deviations of the substations
Regression of the standard deviations over the mean loads of the substations.

global characteristics remain, there are gaps in the behavior between the distribution of the aggregated loads and the distribution of each individual substation.

For instance, we present in Figure 2.18 four substations that behave differently depending on the hour of the week. A similar case for the day of the year is presented in Figure F.17.

More generally, the quantiles over the substations of the conditional loads, normalized by the mean of the substations, are presented in Figure 2.19 for the hour of the week and in the five Figures F.18 - F.22 for the other calendar and meteorological variables.

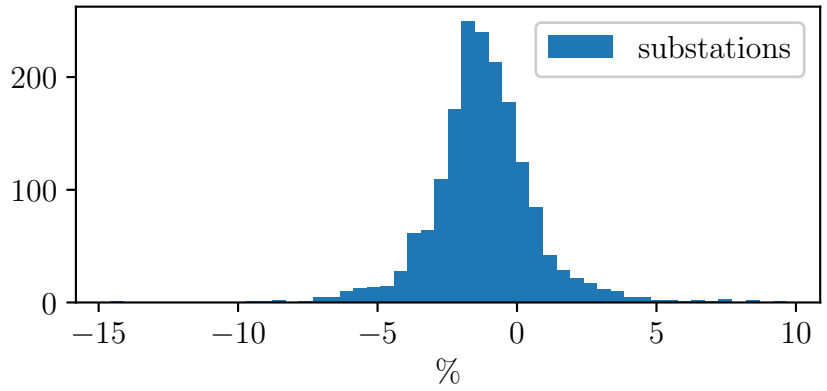


FIGURE 2.16: Regression of the local loads over the timestamp. Distribution of the relative slopes $100 \times \beta_k/\hat{\ell}_k$ for the different substations $k = 1, \ldots, K$ where β_k is the slope obtained by regressing the monthly average load of substation k on the timestamp and $\hat{\ell}_k$ is the average load.

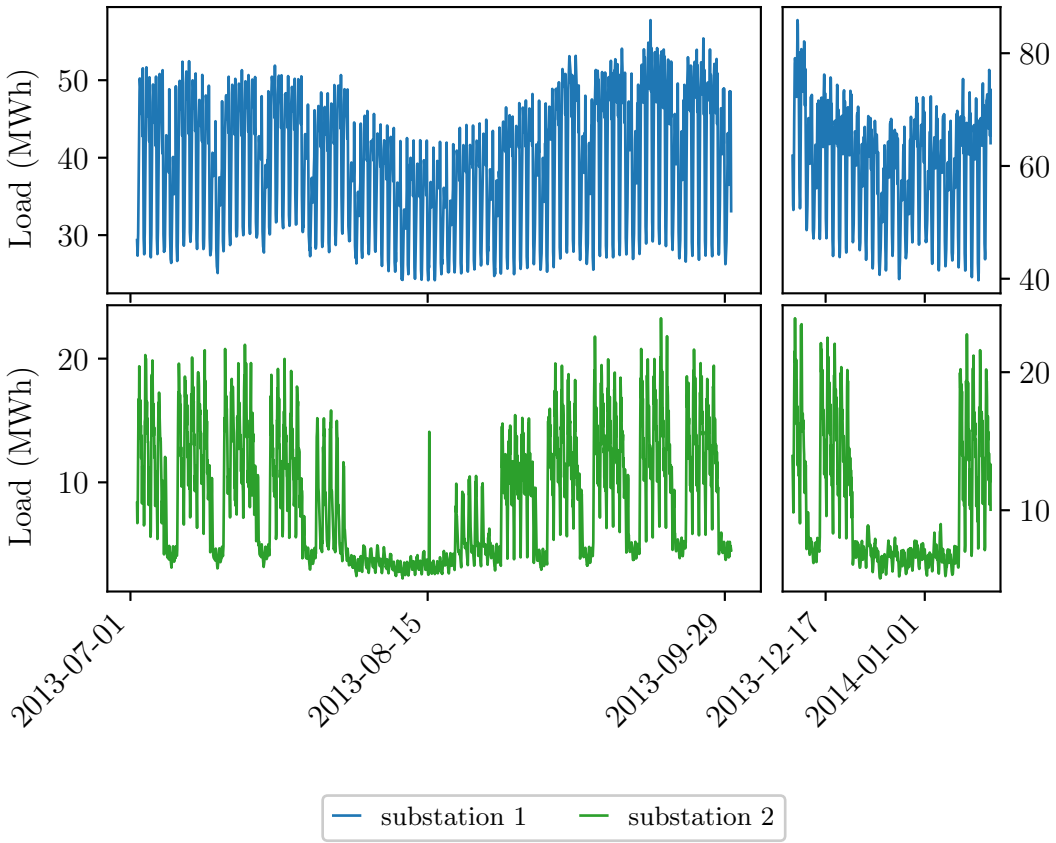


FIGURE 2.17: Local impacts of vacations
Impact of the summer break for two different substations : one substation whose behavior is close to the national load (substation 1) and one that is particularly impacted by the summer and the Christmas periods (substation 2).

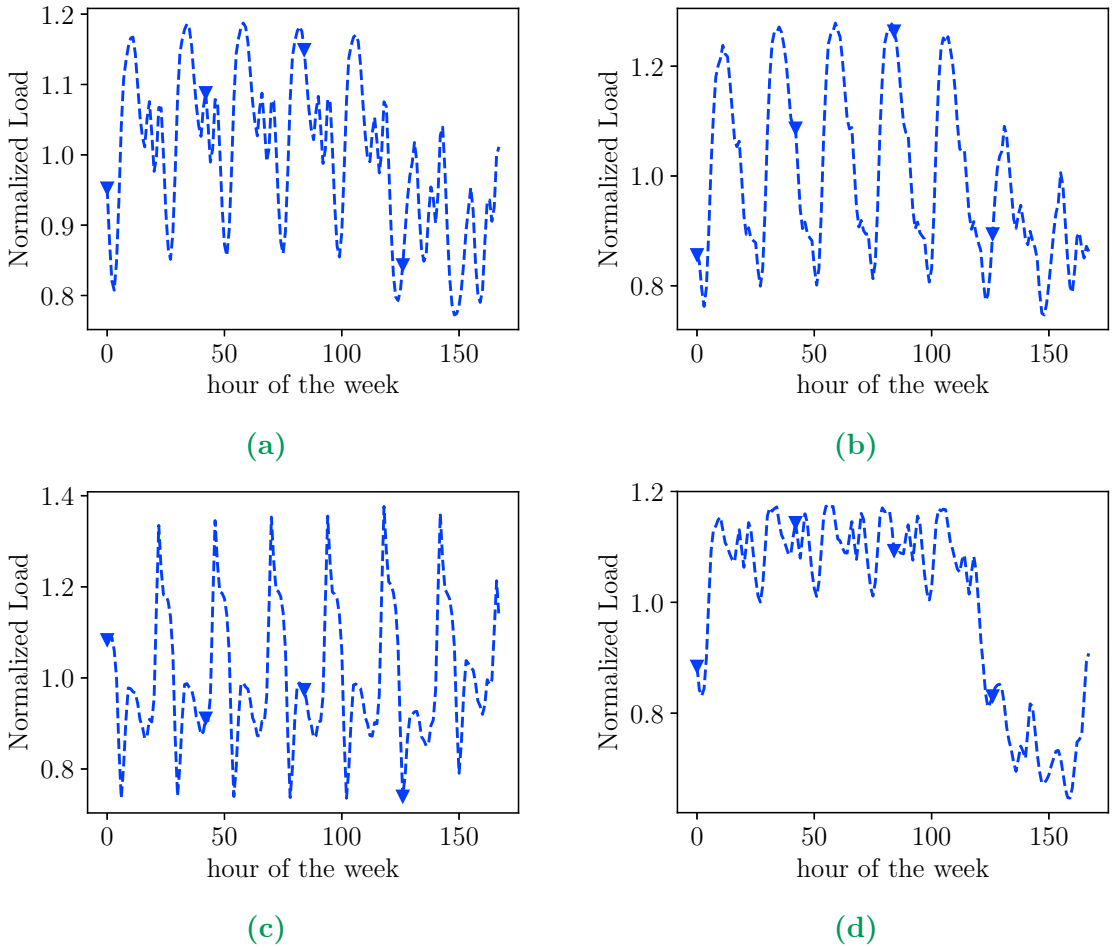


FIGURE 2.18: Diversity of weekly cycles
Empirical expectations of the load divided by its average value, at 4 different substations conditioned on the hour of the week. The normalized load in Figure 2.18a is rather similar to the national load presented in Figure 2.10, the 3 others are quite different. Figure 2.18b presents large peaks in the morning while in Figure 2.18c, the majority of the demand happens in the evening. In Figure 2.18d, the normalized load is significantly diminished during the weekends.

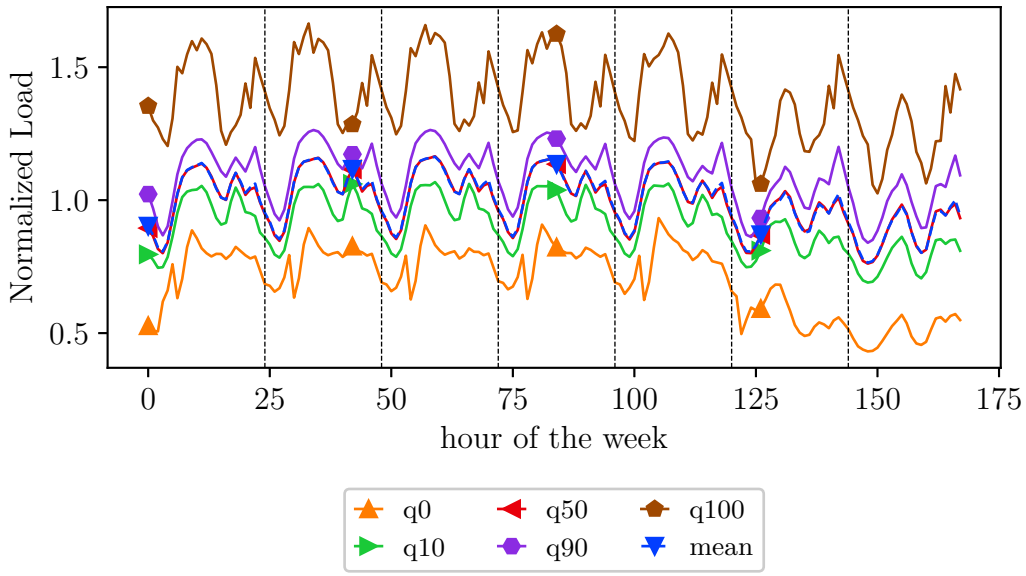


FIGURE 2.19: Local loads conditioned on the hour of the week
Quantiles over the substations of the empirical expectations conditioned on the hour of the week of the centered normalized loads. The vertical lines separate the different days of the week.

2.4.2 Existence of a common structure

Although the presentation in Section 2.4.1 emphasizes the heterogeneity among the substations, there are still strong similarities between the joint distribution of the load and the input variables. They are particularly visible on the empirical conditional expectations of the loads, presented in Section 2.4.1. We call these resemblances the (unidentified) common structure.

Under the condition that this common structure is sufficiently pervading, it might be relevant to mutualize the information at different substations by coupling the individual models, which is the ambition of this work. In this section, we provide additional evidence that supports this idea.

Correlation of the substations Due to the yearly and weekly cycles, we expect the loads of the different substations to be highly correlated. This is indeed confirmed in Figures 2.20 and F.15.

We also present in Figures 2.21 and F.16 the correlations between the substations after subtracting an estimator of the load computed with kernel regression for each day of the year and each hour of the week.

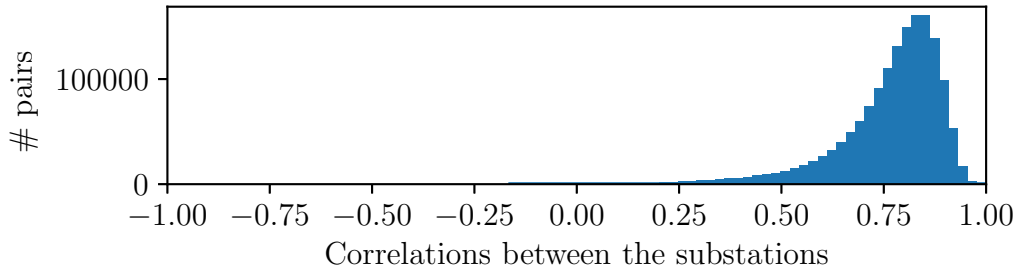


FIGURE 2.20: Correlations between the substations
Histogram of the correlation between the substations over the 5-year-long dataset. Note that we should investigate whether the negative correlations might be due to load reports or have another interpretation.

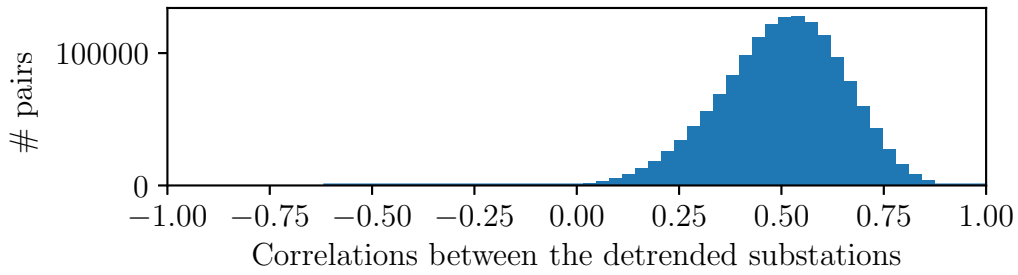


FIGURE 2.21: Correlations between the *detrended* substations
Histogram of the correlations between the substations after an estimate of the load for each day of the year and hour of the week is subtracted. The estimate is obtained with a kernel regression on the observations.

In addition, the matrix of load observations can be relatively well approximated by a low-rank matrix, as illustrated in Figure 2.22. We observe that 90 % of the variance in the columns of the load matrix can be explained by 10 principal components. Thereby, it is legitimate to believe that the necessary complexity to explain the loads at all substations is less than proportional to their number.

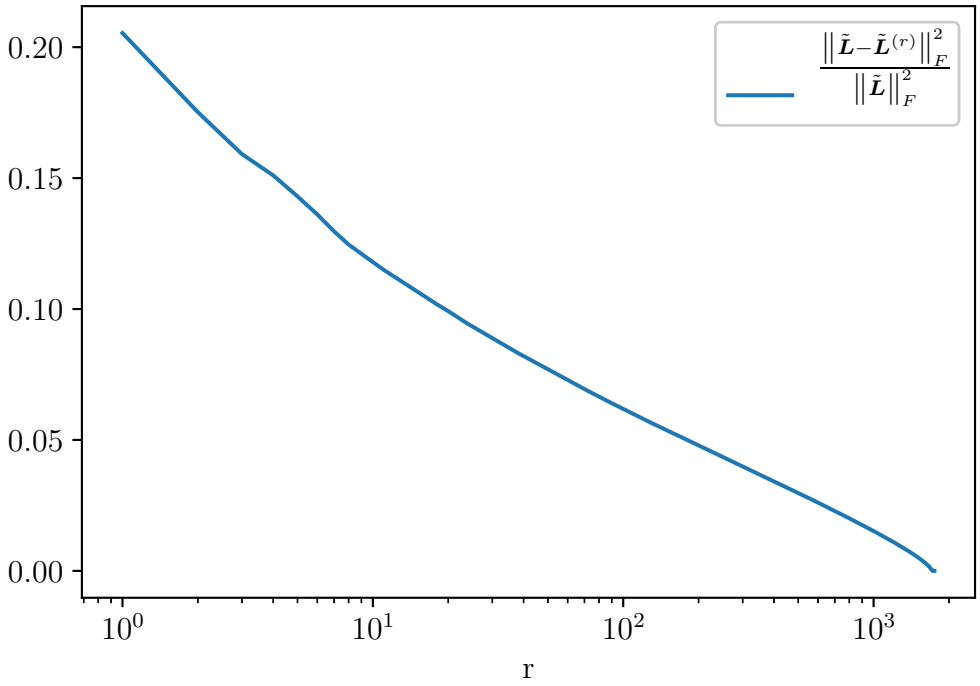


FIGURE 2.22: Singular values of the load matrix
The matrix $\tilde{\boldsymbol{L}}$ is obtained by centering the rows of the original load matrix $\boldsymbol{L} \in \mathbb{R}^{n,\mathcal{K}}$ containing $n = 43824$ observations of the loads over the 5 years in the dataset at the $\mathcal{K} = 1751$ substations. For $r \in \mathbb{N}$, the matrix $\tilde{\boldsymbol{L}}^{(r)}$ is the closest rank-r approximation of $\tilde{\boldsymbol{L}}$ in terms of the Frobenius norm. Note that if $P\text{diag}(s_1, \ldots, s_{\min(n,\mathcal{K})})Q^T$ is a singular value decomposition of $\tilde{\boldsymbol{L}}$, with $P \in \mathbb{R}^{n,\min(n,\mathcal{K})}$, $s_1 \geq \ldots \geq s_{\min(n,\mathcal{K})} \geq 0$ and $Q \in \mathbb{R}^{\mathcal{K},\min(n,\mathcal{K})}$, then the best rank-$r$ approximation of $\tilde{\boldsymbol{L}}$ is the matrix $P\text{diag}(s_1, \ldots, s_r, 0, \ldots, 0)Q^T$ and the plot is simply the graph of $r \mapsto \sum_{t=r+1}^{\min(n,\mathcal{K})} s_t / \sum_{t=1}^{\min(n,\mathcal{K})} s_t$. The value of this function for $r = 0$ equals 1 but the plot begins at $r = 1$ because of the logarithmic scale.

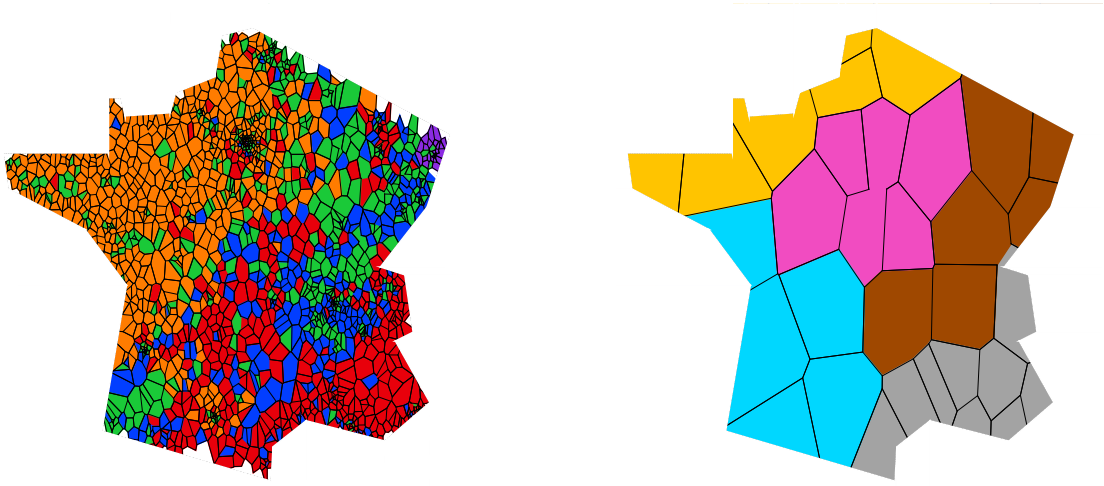
Clustering the substations While the strong correlations and the low-rank approximation presented previously support the claim that the substations have a common underlying structure, they do not give an idea of its organization.

Because the load at different substations have varying amplitudes, using a standard K-means clustering algorithm [MacQueen et al., 1967] to group the resembling substations together is probably not appropriate because there is a risk that the substations are clustered depending on their amplitudes. Instead, it seems more relevant to group the substations depending on their behavior over time and for different values of the inputs. We propose a possible partition of the substations

into 5 clusters in Figure 2.23a, which is obtained with a subspace clustering algorithm [Elhamifar and Vidal, 2013].

Figure 2.23a is visually satisfying. Indeed, without being aware of the geographical coordinates of the substations, the subspace clustering algorithm found groups relatively well organized spatially. However, it is not clear that this clustering is really informative since the substations may have been grouped together only because they are exposed to the same local weather conditions, while we would like to measure more precisely how similarly the loads behave when the distribution of the calendar and the meteorological variables are the same. A subspace clustering of the weather stations into 5 groups is presented in Figure 2.23b and it does not make this doubt disappear.

The interpretations drawn from these visualizations should be taken with caution and measuring the similarity between two substations remains an open problem.



(a) Subspace clustering into 5 groups of the 1751 substations.

(b) Subspace clustering into 5 groups of the 32 weather stations.

FIGURE 2.23: Clusters of substations and weather stations

Forecast from a few leaders Alternatively, we tried to assess how much information is shared between the load curves by performing a linear regression on individual loads with all the others. If every substation were a linear combination of the others, we could obtain a perfect estimation of the load at one substation from all the others. A simple extension of this idea leads to wonder whether we could in fact predict all the substations with a few of them, that we call the leaders. Such a task can be performed by simultaneously learning all the regression models, adding a group-Lasso regularization term and solving the following optimization problem :

$$\min_{\boldsymbol{C}\in\mathbb{R}^{\mathcal{K},\mathcal{K}}} \frac{1}{2n} \|\boldsymbol{L} - \boldsymbol{L}\boldsymbol{C}\|_F^2 + \lambda \|\boldsymbol{C}\|_{1,2} \,, \tag{2.2}$$

where $\lambda \geq 0$, $\boldsymbol{L} \in \mathbb{R}^{n,\mathcal{K}}$ is the load matrix with n observations at $\mathcal{K}$ substations, $\boldsymbol{C} := (c_\ell^k)$ is the coefficient matrix to learn and :

$$\|\boldsymbol{C}\|_{1,2} := \sum_{\ell=1}^{\mathcal{K}} \sqrt{\sum_{k=1}^{\mathcal{K}} (c_\ell^k)^2} = \sum_{\ell=1}^{\mathcal{K}} \|\boldsymbol{c}_\ell\|_2 \,, \tag{2.3}$$

with $\boldsymbol{c}_\ell$ the ℓ-th row of the matrix $\boldsymbol{C}$. The group-Lasso regularization [Bakin et al., 1999; Obozinski et al., 2010; Yuan and Lin, 2006] is known for encouraging some of the groups to have zero norm. Thereby, it induces in our case some of the rows of $\boldsymbol{C}$ to be zero, which eliminates them from the set of predictors. Note that these forecasts are not feasible in practice since we need the loads of the leaders for the instants to forecast.

Because the group-Lasso regularization induces a bias and due to the nice properties of the Elastic-Net penalty as a variable selection procedure [Zou and Hastie, 2005], we instead proceeded sequentially as follows. First, we solve the selection problem with the Elastic-Net penalty :

$$\min_{\boldsymbol{C} \in \mathbb{R}^{\mathcal{K},\mathcal{K}}} \frac{1}{2n} \|\boldsymbol{L} - \boldsymbol{L}\boldsymbol{C}\|_F^2 + \lambda \left(\alpha \|\boldsymbol{C}\|_{1,2} + \frac{(1-\alpha)}{2} \|\boldsymbol{C}\|_F^2 \right), \tag{2.4}$$

where $\lambda > 0$ and $\alpha = 0.99$. Let $\check{\boldsymbol{C}}$ denote the concatenation of the $r \in \mathbb{N}$ non-zero rows of the estimated solution of Problem (2.4) and $\check{\boldsymbol{L}}$ the matrix obtained by concatenating the selected columns. The columns of $\check{\boldsymbol{L}}$ are considered as a new set of predictors and we then solve :

$$\min_{\boldsymbol{D} \in \mathbb{R}^{r,\mathcal{K}}} \frac{1}{2n} \left\|\boldsymbol{L} - \check{\boldsymbol{L}}\boldsymbol{D}\right\|_F^2 + \frac{\mu}{2} \|\boldsymbol{D}\|_F^2 \,. \tag{2.5}$$

The results of this 2-step procedure are presented in Figure 2.24. With $\lambda = 1$ in the first step, 14 substations are selected and using them as a set of predictors leads to an average $\mathtt{r}^2$ over the substations of 0.8. With $\lambda = 0.1$, 100 substations are selected and they lead to an average $\mathtt{r}^2$ of 0.9. This confirms that the loads in the different substations have a common structure.

Outliers Although we provided evidence that there is an underlying structure common to a significant number of the substations, we must keep in mind that there are outliers : some substations have their own particular behavior and it would probably not be a good idea to mutualize parts of the learning process with them. We allowed ourselves to discard a significant amount of the substations, 338 out of 2089, from the database with the correction procedure described in Section B.2. Still, there may remain outliers that cannot be well forecast with the selected leaders, as illustrated in Figure 2.25, and we do not have a clear criterion to know which ones to discard yet.

As a conclusion of this section, there is heterogeneity among the substations, but it is reasonable to believe that the data is somehow structured. The whole point of this work is to try to leverage this unknown structure to improve the quality of the forecasts.

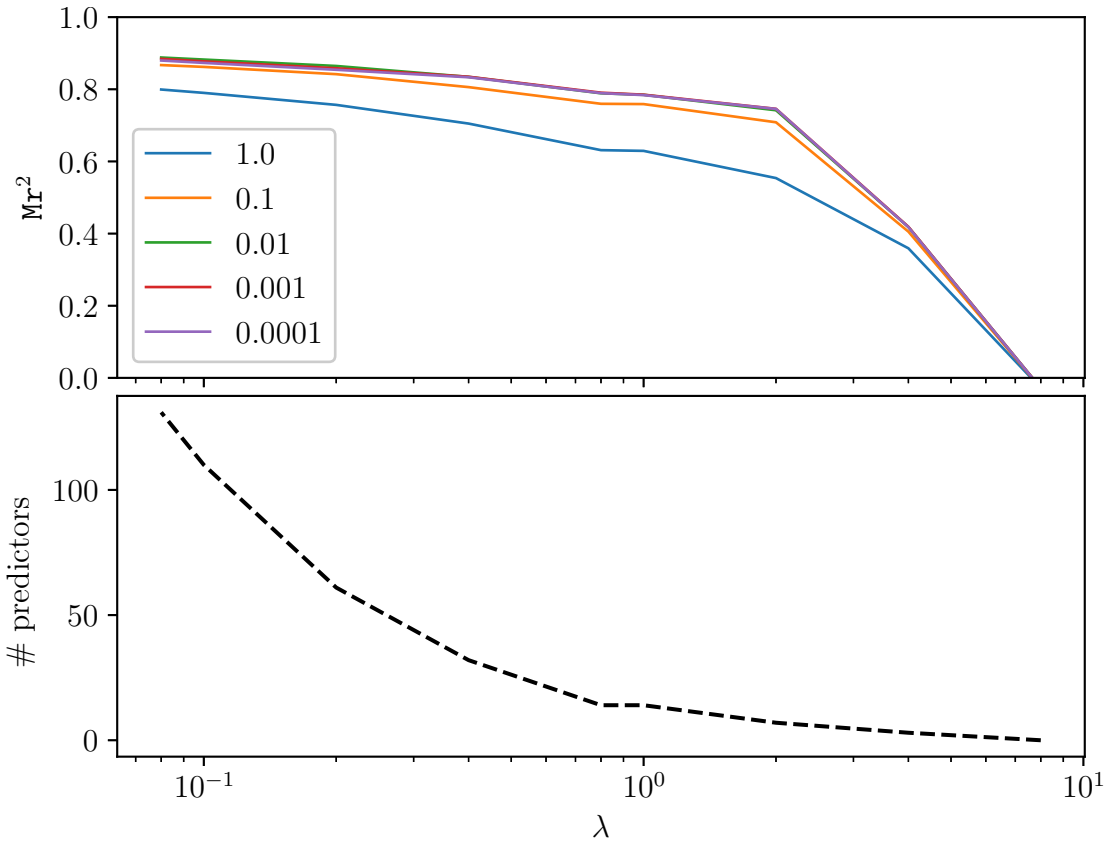


FIGURE 2.24: Forecasts with a few leaders

(*top*) Average coefficient of determination $\mathtt{Mr}^2$ over all the sub-stations of the individual coefficients of determination $(\mathtt{r}_k^2)_{k\in[\![1,\mathcal{K}]\!]}$ for different values of λ, each curve corresponding to a different values of the hyperparameter μ in the second step.

(*bottom*) Numbers of predictor selected in the first step for different values of λ.

In this procedure, we have a 3-year-long training set, from 2013 to 2015 and the performances of the model are computed in 2016.

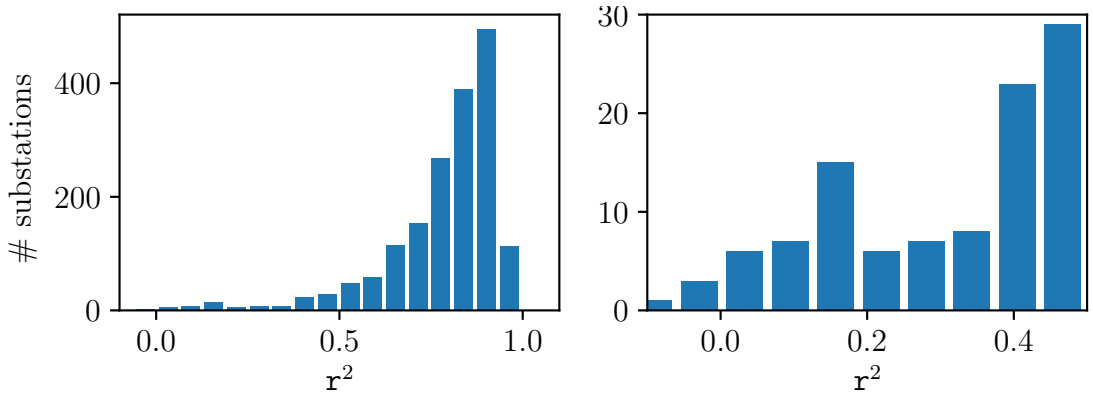


FIGURE 2.25: Presence of outliers
(*left*) Histogram of the coefficients of determination $(\mathtt{r}_k^2)_{k\in[\![1,\mathcal{K}]\!]}$.
(*right*) Same histogram zoomed on the outliers.
The hyperparameters for the 2-step procedure described by Equation (2.4) and Equation (2.5) are $\lambda = 1$ so that 14 leaders are selected and $\mu = 0.001$. Other than that, the learning process is the same as for Figure 2.24.

2.5 Problem settings

Having presented the database and the behaviors of the different time series, we now describe more precisely the forecasting problems considered in this work. A load forecasting problem is defined by a perimeter, a time horizon and the available information. In this manuscript, we are only interested in day-ahead load forecasting but we distinguish 2 possible settings for the available data and 5 different aggregation levels.

2.5.1 Middle-term and short-term models

The shorter the time horizon, the more recent information is available to forecast electricity loads. While more settings can be considered, we only introduce two problems that seem particularly relevant for the TSO. In short, they differ on the availability of the recent loads among the inputs and have consequently distinct use cases.

Middle-term forecasts The middle-term model is only based on exogenous inputs : the forecasts depend on the calendar variables and the meteorological conditions but do not depend on the past loads. With the adequate weather scenarios, a middle-term model can reasonably be used in practice to forecast the load several days or weeks ahead, the main limitation being the accuracy of weather forecasts for long time horizons. Nevertheless, a middle-term model can also serve as an estimator of the average load during this period if average weather conditions are available for a future time period of the year. More generally, it provides a simple tool to analyze the relationships between the weather, the time and the loads.

Short-term forecasts In addition to the information available to a middle-term model, a short-term model also has access to endogenous information through the recent values of the target time series. It constitutes a significant advantage over the middle-term model and clearly impacts the performances.

While the relationships of the loads with the weather or the economic activity are relatively well-understood, the introduction of the past loads in the inputs has a less clear interpretation. Certainly, it provides information on the economic activity that are not contained in the calendar data but we could equally conjecture that it informs the model of a sensitivity to other weather conditions. Improvements with the short-term models could also indicate that the expressiveness of the middle-term model is inappropriately limited. We do not go further in this interpretation and use the model with the best performances, considering that the past loads provide complementary information.

2.5.2 Aggregation levels

Aggregated Loads Because the priority of the TSO is to ensure the supply and demand equilibrium, the first load forecasting problem to be addressed was for the national demand. A few years ago, this work was extended to forecast the load of the administrative regions and more recently, the load forecasting at the substations

level was considered. In this manuscript, we are especially interested in the latter setting but introduce two more intermediate aggregation levels, respectively defined by the organization of the network and by the position of the weather stations. Their goal is to enrich the set of possible problems.

An aggregation level of the $\mathcal{K}$ substations in the database is characterized by a partition $\mathcal{Z} := (Z_k)_{k\in[\![1,K]\!]}$ of the $\mathcal{K}$ substations into K zones. Given $\kappa \in [\![1, \mathcal{K}]\!]$, we denote r_κ the load of substation κ and for $k \in [\![1, K]\!]$, we denote :

$$\ell_k := \sum_{\kappa\in Z_k} \mathsf{r}_\kappa, \tag{2.6}$$

the sum of the loads in zone Z_k. For simplicity and because we do not consider different aggregation levels simultaneously, we omit the partition $\mathcal{Z}$ in the notation.

Weather information Given a zone Z_k in a partition $\mathcal{Z}$, the weather information extracted from the $\mathcal{W} = 32$ weather stations and injected in the modeling of the electricity load of Z_k should obviously depend on Z_k. We consider two possibilities.

First, to model the load in the zone Z_k, we can consider a linear combination of all the weather stations with weights $\boldsymbol{\alpha} \in [0, 1]^{\mathcal{W}}$ such that $\sum_{s=1}^{\mathcal{W}} \alpha_s = 1$, we denote the mean of the temperatures weighted by the vector $\boldsymbol{\alpha}$:

$$\mathsf{T}^{\boldsymbol{\alpha}} := \sum_{s=1}^{\mathcal{W}} \alpha_s \mathsf{T}^s, \tag{2.7}$$

where T^s is the temperature at the weather station $s \in [\![1, \mathcal{W}]\!]$. The same notation is used for the cloud covers :

$$\mathsf{c}^{\boldsymbol{\alpha}} := \sum_{s=1}^{\mathcal{W}} \alpha_s \mathsf{c}^s. \tag{2.8}$$

Such linear combinations are used for instance for the operational forecasting model of the national load.

In the second case, the zone Z_k is associated with a subset $\mathcal{W}_k \subset [\![1, \mathcal{W}]\!]$ of the weather stations and the forecasts of the aggregated load in zone Z_k only rely on the weather information extracted from the weather stations in $\mathcal{W}_k$, the temperatures at the different substations being injected in the models as distinct inputs.

National setting In the national setting, the partition is made of a single set that contains all the substations and the goal is to forecast the sum of all the loads :

$$\ell_{\text{national}} := \sum_{\kappa=1}^{\mathcal{K}} \mathsf{r}_\kappa \tag{2.9}$$

In the historical national model, RTE has decided of the linear combination given in Table F.1 of the 32 weather stations [RTE, 2011]. Thereby, a single fictive temperature obtained as a weighted average of the 32 weather stations is used for the national setting.

Substations level The local forecasting problem corresponds to the prediction of the loads $(\mathsf{r}_\kappa)_{\kappa=1,\dots,\mathcal{K}}$, at each individual substation illustrated in Figure 2.26. Consequently, we set for the substations level $K = \mathcal{K}$ and for each $k \in [\![1, \mathcal{K}]\!]$, $\ell_k := \mathsf{r}_k$. It is not relevant to consider for each substation all the weather stations or the previously defined weighted mean of the weather stations. Instead, we consider for a substation $k \in [\![1, K]\!]$ the two weather stations $s_1^k, s_2^k \in [\![1, \mathcal{W}]\!]$ that are geographically closest to the substation k. This decision is discussed in Section 3.6.5.

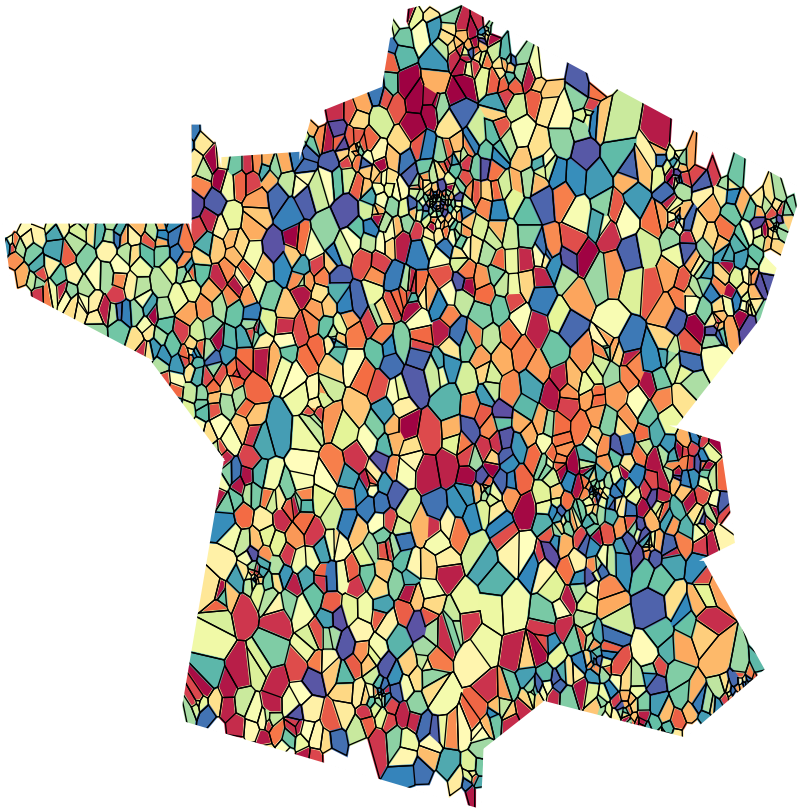


FIGURE 2.26: Voronoi diagram of the substations
Note that the size of each area is not proportional to the load. Large cities correspond to regions with a high density of substations that have consequently small areas on the map.

The time series at the level of the substations are much noisier. They may also present heterogeneous behaviors that make the local load forecasting problem significantly different from the national problem. For this reason, we introduce intermediary settings.

RTE regions Based on the topology of the high-voltage network, the TSO partitioned the country in 7 regions that we denote $N_1, \dots, N_7$. In order to forecast the aggregated loads in one of these regions $k \in [\![1, 7]\!]$, we use the two weather stations

that are closest to the center of the region. Thus we obtain 7 aggregated loads $(\ell_k)_{k=1,\dots,7}$ as presented in Figure 2.27 and Table 2.1.

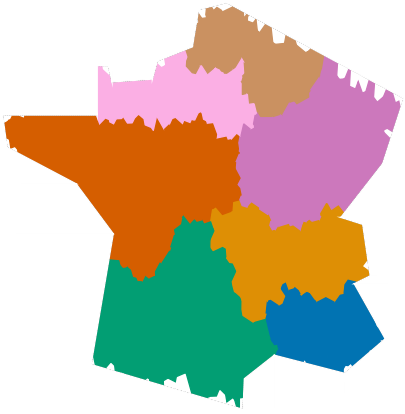


FIGURE 2.27: Map of the 7 RTE regions
Map of the 7 Regions defined by the high-voltage network. First we have computed the Voronoi diagram of the substations and secondly painted each area with the color of the corresponding regions.

Administrative regions The 12 administrative metropolitan regions of France that we denote $A_1, \dots, A_{12}$ form a slightly thinner partition of the country and have consequently slightly noisier load time series $(\ell_k)_{k=1,\dots,12}$. They are described in Figure 2.28 and Table 2.1. The weather stations associated to each administrative region are also the two closest to the center of the region.

FIGURE 2.28: Map of the 12 metropolitan administrative regions

Districts We introduce 32 districts $D_1, \dots, D_{32}$ that form a partition of the whole set of substations, defined by the Voronoi diagram of the weather stations presented

in Figure 2.29. As explained in Table 2.1, this setting lies between the administrative setting and the local setting. Although the corresponding time series $(\ell_k)_{k=1,\dots,32}$ are more sensitive to the local weather conditions than for the coarser regions, noise is reduced compared with the loads of the individual substations. To simplify, we consider that each district only has access to the temperature of the associated weather station.

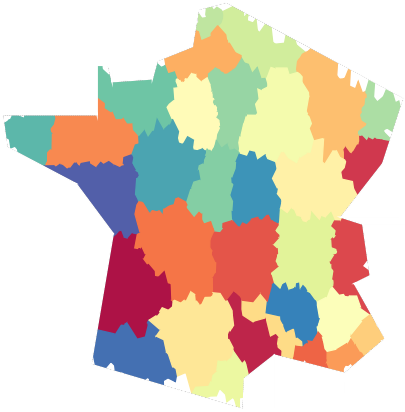


FIGURE 2.29: Map of the 32 districts

	National	RTE regions	Administrative regions	Districts	Local
K	1	7	12	32	1751
$\mathcal{K}/K$	1751	250	146	55	1
$\bar{\ell}$ (MWh)	39 000	5 500	3 200	1 200	21
$\lvert S\rvert$	1	2	2	1	2

TABLE 2.1: Characteristics of the different aggregation levels
The number of zones for an aggregation level is denoted K and the average number of substations per zone is $\mathcal{K}/K$. The average hourly load of the zones is denoted $\bar{\ell}$ and the number of weather stations that we use to model the load in each zone within an aggregation level is denoted $|S|$. Note that the unique station used at the national level is fictive and obtained with the linear combination of Equations (2.7) and (2.8). Besides, the choice of the number of weather stations at the local level is discussed in Section 3.6.5.

2.6 Related work - Load forecasting

The modeling of the relationship between the electricity demand, the calendar information and the weather conditions has been the object of interest of both the statistical and the economy communities for the last few decades. The models based on socio-economic information and the analyses of electricity end-uses have proved more relevant for long-term horizons, of several months or years while the Machine Learning and more generally the statistical modeling, sustained by the increase in computing power of modern machines and by the vast amount of collected data, have dominated among the approaches to load forecasting problems with short-term horizons.

Broad surveys on the topic of load forecasting include [Hahn et al., 2009; Kyriakides and Polycarpou, 2007; Muñoz et al., 2010; Weron, 2007] and the models presented often extend to other quantities of current interest for forecasting and power systems including energy prices [Nowotarski and Weron, 2018] and renewable energy production, which establish a connection with short-term weather forecasts techniques [Cros and Pinson, 2018; Messner and Pinson, 2018; Nagbe et al., 2017; Petra et al., 2014].

Note that in operational conditions, most forecasting tools are different from the models encountered in the literature because the forecasts are often manually modified *a posteriori* by forecasters, in particular to take into account special and punctual events having a noticeable impact on the electricity demand.

Statistical approach A wide variety of forecasting methods have been proposed to model the electricity load, because no model has proved to be significantly better than the others in all possible settings, even for the short-term forecasting of the aggregated load at regional or national levels. Pioneer works on electricity load forecasting applied classical statistical tools, notably autoregressive models. Their flexibility and especially their ability to include seasonal components, trends and effects of exogenous variables [Huang and Shih, 2003; Nowicka-Zagrajek and Weron, 2002] has justified their use as classical benchmarks for load forecasting problems. Exponential smoothing techniques have equally been considered [Taylor, 2010, 2011].

Multilayer prediction The complex relationships between the input variables and the electricity demand lead researchers to consider more sophisticated models. Thereby, the universal approximation of neural networks motivated non-statistical modeling [Hippert et al., 2001; Khotanzad et al., 1997; Kiartzis et al., 1995; Park et al., 1991]. A significant improvement was finally obtained in 2004 with Support Vector Machines [Chen et al., 2004].

The potential of tree-based models able to model high non-linearities with weak learners at a low computational cost was also assessed by Dudek [2015]. Additionally, the design and the aggregation of specialized load forecasting experts was studied by Devaine et al. [2013]; Gaillard and Goude [2015], with models estimated over different time windows by Pesaran and Pick [2011], and with a procedure of selection for high-dimensional data modeled with functional regression by Mougeot et al. [2015].

Generalized Additive Models From 2011, the successful application to the load forecasting problem and the interpretability of the Generalized Additive Models (**GAM**) based on the calendar variables, the weather and the past values of the series has motivated a deeper analysis and various extensions[2]. In particular, they lead to an improvement of the predictions compared with the historical additive model used by EDF [Bruhns et al., 2005], that requires expert knowledge to be tuned and is considered to be insufficiently modular. That is why they are given a particular attention in this manuscript and are discussed in more details in Section 2.9.4. In particular, Pierrot and Goude [2011] specialized these models to the modeling of the French national demand and Goude et al. [2013] pursued this approach to model the electrical load of about 2000 substations of the French distribution network.

Variable selection For an adaptation to high-dimensional inputs and outputs, the GAM were also considered simultaneously with a 2-step variable selection procedure [Thouvenot, 2015; Thouvenot et al., 2015], first with a selection of the relevant inputs variables with a group-Lasso regularization like in Equation (2.3) and a tuning of the regularization hyperparameters based on a Model Selection Criteria [Akaike, 1974; Craven and Wahba, 1978; Shenoy et al., 2015], then with a relaxed version of the objective, *i.e.* without the group-Lasso regularization, to correct the bias induced by the latter [Zhang et al., 2008]. In addition, Thouvenot et al. [2015] provided a statistical analysis of their estimator and proved its consistency for variable selection.

State-Space Models As a major shortcoming of the aforementioned models, the difficulty to model the non-stationarity of electricity demand was addressed with periodic State-Space Models (**SSM**), able to adapt to changes of regime and long-term non-stationarity of the electricity consumption [Dordonnat et al., 2008]. A functional vector autoregressive SSM based only on endogenous data was proposed by Nagbe et al. [2018].

Modeling uncertainty More recently quantile regression and density forecasting, that is to say the prediction of a whole conditional distribution, were the objects of an increasing attention of both the Machine Learning community [Dawid, 1984; Sangnier et al., 2016] and the users of the load forecasting models [Hong and Fan, 2016; Shenoy et al., 2015] as well as the wind power forecasting models [Pinson, 2012; Sloughter et al., 2010].

With stochastic process modeling and the estimation of a confidence interval, Antoniadis et al. [2014] extends the work of [Antoniadis et al., 2012], whose general principle consists in finding in the history, observations similar to the present-day context in order to provide a forecast based on a linear combination of the similar observations, where the similarity is measured with the coefficients obtained by a

[2]In the electricity load forecasting literature, these models are sometimes called semi-parametric models. This denomination seems less appropriate than GAM since what matters most is their nonlinearity, which leads to the adjective Generalized, and their Additive structure, but not their potential infinite parametrization. Besides, it is very rare to have a load forecasting model that actually has an infinite number of parameters to estimate, mainly because of the Representer Theorem [Wahba, 1990, and references therein].

Kernel Wavelet transform [Antoniadis et al., 2006]. A comparable approach based on curve linear regression to forecast a day of consumption given its recent past was developed by Cho et al. [2013, 2015].

Alternatively, bootstrapping methods were considered almost 10 years ago [Fan and Hyndman, 2011] and more recently with randomly generated temperature scenarios [Gaillard et al., 2016]. Meanwhile, Gaillard et al. [2016] approached the problem with the pinball loss developed for quantile regression [Koenker, 2005; Koenker and Bassett Jr, 1978]. Finally, an estimation of the time-varying covariance matrix in GAM was studied by Wijaya et al. [2015].

Particularly relevant and studied in meteorology, the problem of choosing an appropriate method to assess and compare empirically density forecasts was addressed by Gneiting et al. [2007] who propose a study of the probability integral transform histogram, marginal calibration plots, the sharpness diagram and proper scoring rules.

Multiple output forecasting A large majority of the load forecasting models presented so far is focused on the load aggregated at regional or national levels. Still, forecasts at non-aggregated levels were considered for buildings or residential neighborhoods [Kolter and Ferreira, 2011; Wijaya, 2015], homogeneous groups of consumers [Cugliari et al., 2016; Wijaya et al., 2014], and geographical areas [Hong et al., 2014]. Additionally, Thouvenot [2015] studies the local load forecasting problem for 61 of the 1751 substations that we consider in this manuscript, in a region near Lyon and with a particular attention paid to the selection of relevant input variables.

Depending on the residential, commercial or industrial nature of the electricity uses contained in these disaggregated time series, the curves may have strong similarities and share an underlying structure. Leveraging such a structure in the modeling to obtain a better generalization performance is the question of interest in Multi-task Learning. The tools developed in this branch of Mathematics have been applied in the last decade to the forecasting of electricity production from renewable sources [Sanandaji et al., 2015; Wytock and Kolter, 2013] and to local loads forecasting problems.

Relying on the hierarchical organization of the time series, Auder et al. [2018] studied the individual (household level) and aggregated (national level) load curves to propose a clustering tool with the same wavelet-based notion of similarity as in [Antoniadis et al., 2012]. Instead, Hyndman et al. [2011] introduced a model where the time series of different levels are forecast independently and then optimally combined with a linear regression model consistently with the hierarchical organization of the network.

Alternatively, Kim and Giannakis [2013] consider low-rank formulation of multi-task load forecasting problems in an attempt to leverage and potentially reveal the underlying structure of the load curves. Promoting the interpretability of non-negative matrix factorization formulation [Lee and Seung, 1999, 2001], Mei et al. [2017] studied the problem of time series recovery in the context of incomplete measurements and extended the model with side-information to times series prediction [Mei et al., 2018].

The rising interest for local load curves has additionally motivated the develop-

ment of methods for the detection of anomalies [Jian et al., 2018, and references therein], much more present at disaggregated levels.

Future stakes Major challenges have recently emerged in the electricity sector, such as the adaptation to modern energy markets, the integration of renewable energies and the penetration of electric vehicles, progressively being reflected in the research literature. The installation of smart meters and the conditions necessary to the realization of their potential progressively draw attention too, to take into account the Demand-Response Mechanisms as well as to leverage the considerable datasets collected, certainly leading to Big Data considerations. Mei et al. [2016] studied for instance the relationship between socio-demographic characteristics and local electricity uses in order to extrapolate the demand in regions with socio-demographic information but few measurements of the electricity demand with smart meters.

2.7 Numerical evaluation

The most relevant evaluation of a model depends on the final task it is designed for. Generally speaking, there is never an easy way to order forecasting methods, especially models with non-stationary multivariate outputs. However, we chose to assess the quality of the models with numerical criteria presented in this section as guides in our research, in order to summarize in a unique number the discrepancy between observed target variables and the predictions.

Note that there might be a post-processing of the forecasts by RTE, in particular when rare events occur. Besides, the day-ahead predictions of the models are always corrected during the intraday forecasting process, once more recent observations are available. None of these two post-processings are considered in this manuscript.

2.7.1 Evaluation criteria for a single-task problem

In a single-task setting, the ℓ^1 and ℓ^2 distances usually provide relevant means to evaluate a forecasting model. Consider a zone Z_k that corresponds to a subset of the substations as introduced in Section 2.5.2, a time period denoted :

$$\mathcal{T}_b := [\![1, n]\!], \tag{2.10}$$

indexed by $b \in \mathbb{N}$ and a batch of observations $\boldsymbol{\ell}^{(k)} := (\ell_i^k)_{i\in\mathcal{T}_b} \in \mathbb{R}^n$. To assess the quality of a prediction $\hat{\boldsymbol{\ell}}^{(k)} := (\hat{\ell}_i^k)_{i\in\mathcal{T}_b}$, we introduce the following performance criteria.

Mean Squared Error The Mean Squared Error (**MSE**) is a simple quadratic penalization of the residuals. It corresponds to the squared Euclidean distance between the forecasts and the observations. It also has the notorious advantage of being a convex and differentiable loss. It is defined as :

$$\texttt{MSE}_{k,b} = \frac{1}{n}\sum_{i=1}^{n}(\ell_i^k - \hat{\ell}_i^k)^2. \tag{2.11}$$

The coefficient of determination $\mathtt{r}^2$ The coefficient of determination $\mathtt{r}^2$ is a relative error that makes the comparison of different tasks more convenient. It is obtained with an affine transformation of the MSE based on the empirical variance of the time series $(\ell_i^k)_{i=1,\dots,n}$. It is without unit and in particular invariant to affine transformations of the target variables. Let

$$\bar{\ell}_k := \frac{1}{n}\sum_{i=1}^{n} \ell_i^k$$

denote an estimate of the average load for the considered area :

$$\mathtt{r}^2{}_{k,b} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(\ell_i^k - \hat{\ell}_i^k)^2}{\frac{1}{n}\sum_{i=1}^{n}(\ell_i^k - \bar{\ell}_k)^2}. \tag{2.12}$$

We always have $\mathtt{r}^2{}_{k,b} \leq 1$ and a negative score means that the predictions are worse than the constant prediction $\bar{\ell}_k$.

Normalized Mean Squared Error Proportional to the MSE, the Normalized Mean Squared Error (**NMSE**) is defined as :

$$\mathtt{NMSE}_{k,b} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\ell_i^k - \hat{\ell}_i^k)^2}{\bar{\ell}_k^2}. \tag{2.13}$$

Related to the NMSE, we also define the Root Normalized Mean Squared Error (**RNMSE**) :

$$\mathtt{RNMSE}_{k,b} = 100 \times \sqrt{\mathtt{NMSE}_{k,b}}. \tag{2.14}$$

MAPE Finally, the Mean Absolute Percentage Error (**MAPE**) is related to the ℓ^1-distance between the observations and the predictions, thereby it is more robust (less sensitive) to outliers. It is defined for time series without any zero value, which is the case of well collected loads :

$$\mathtt{MAPE}_{k,b} = 100 \times \frac{1}{n}\sum_{i=1}^{n} \frac{|\ell_i^k - \hat{\ell}_i^k|}{|\ell_i^k|}. \tag{2.15}$$

Comments In the single-task setting, the MSE, the coefficient of determination and the NMSE are affine transformations of each other. They are commonly used as loss functions, in large part because they are $\mathcal{C}^\infty$ functions. However, as quadratic functions, they are sensitive to outliers while the MAPE is a more robust criteria.

In practice, we minimize in this manuscript the NMSE and mainly compare single-task models using the RNMSE. Also, note that the coefficient of determination and the NMSE are highly sensitive to the empirical mean and variance of the time series. In particular, computing these criteria for a prediction of one year is not equivalent to averaging the same criteria computed separately for each of the 12 months : one should always compare performances for test sets that correspond exactly to the same period since the mean and the variance of the electrical load are generally twice larger during the winter than during the summer.

2.7.2 Evaluation criteria for the multi-task setting

We now discuss how to evaluate a forecasting model on several possibly heterogeneous time series. We consider $K \in \mathbb{N}^*$ tasks that correspond to different substations or areas of consumption.

We require the numerical criteria to be representative of the cost induced by the errors in the model for the TSO. In particular, the TSO asked that the numerical criteria should reflect the following : an error of x MWh of the prediction for a small area is more damaging to the network management than an error of the same amplitude in a larger area. Therefore, we introduce, for a time period $\mathcal{T}_b$, the Mean NMSE (**MNMSE**) and the Root MNMSE (**RMNMSE**) :

$$\texttt{MNMSE}_b = \frac{1}{K}\sum_{k=1}^{K} \texttt{NMSE}_{k,b}, \tag{2.16}$$

$$\texttt{RMNMSE}_b = 100\sqrt{\texttt{MNMSE}_b}. \tag{2.17}$$

The MNMSE is the criteria used for optimization in the multi-task problems cast in this manuscript. The first mean is taken in Equation 2.13 with respect to the observation instants and the second mean in Equation 2.16 with respect to the different tasks. Our results presents the RMNMSE along with the average coefficient of determination and the average MAPE :

$$\texttt{Mr}^2_b = \frac{1}{K}\sum_{k=1}^{K} \texttt{r}^2_{k,b}, \tag{2.18}$$

$$\texttt{MMAPE}_b = \frac{1}{K}\sum_{k=1}^{K} \texttt{MAPE}_{k,b}. \tag{2.19}$$

We could go even further and require that the criteria also reflects that an error of x % in a large area is more damaging than an error of x % in a smaller area of consumption. For this reason, we also introduce the Weighted RNMSE and the weighted MAPE :

$$\texttt{WRNMSE}_b = \frac{\sum_{k=1}^{K} \bar{\ell}_k \times \texttt{RNMSE}_{k,b}}{\sum_{k=1}^{K} \bar{\ell}_k}, \tag{2.20}$$

$$\texttt{WMAPE}_b = \frac{\sum_{k=1}^{K} \bar{\ell}_k \times \texttt{MAPE}_{k,b}}{\sum_{k=1}^{K} \bar{\ell}_k}. \tag{2.21}$$

It is not clear which of these criteria is the most relevant, industrially speaking. Eventually, we mainly focus on the RMNMSE.

To compute a unique quantity from the performances of the model for different tasks, we chose to compute a uniform or weighted mean over the tasks in the above formulas. Alternatively, we could compute the median, it has the notable advantage of not being as sensitive to outliers as the mean and in particular, not being as sensitive to the choice of the discarded substations with the correction procedure of Section B.2. Still, we use the average and illustrate the distribution of the performances of the different tasks when relevant.

2.7.3 Experimental process

In operational conditions, the forecasts are made day by day and we are allowed to modify the model everyday by incorporating in the training data the most recent days that were already forecast. However, updating the model everyday and forecasting the day in the test set one by one is tedious in the experiments and not necessarily relevant : the fitted model does not change drastically if we add in the training data only the last observed day. Instead, we choose to update the model every s observations with $s \in \mathbb{N}^*$, and perform repeated experiments with sliding training and test sets. Empirically, we estimated that updating the models every 4 weeks is reasonable. This is discussed in more details in Section 3.7.4.

Additionally, because of the obsolescence of data, it is not true that the bigger the training set, the better the fitted model. For instance, we estimated empirically that training sets with 3 years of data are reasonable for the national forecasting problem and we also discuss this decision in Section 3.7.4. As an additional remark, we also observed that the optimal length of the training sets may vary between years in the database. We denote h the length of the training sets and consider it as a hyperparameter of the models.

In summary, the models are repeatedly learned with sliding training sets of size h and evaluated on smaller test sets of size s. More formally, let $[\![L^{\text{test}}, R^{\text{test}}]\!]$ denote an interval in the database corresponding to the whole period used to test the models. We consider that older observation instants $[\![L^{\text{test}} - h, L^{\text{test}} - 1]\!]$ are available for training. For simplicity, we also assume that there exists $B \in \mathbb{N}^*$ such that $R^{\text{test}} - L^{\text{test}} + 1 = B \times s$. For $b \in [\![0, B-1]\!]$, we train the models with the b-th training set :

$$\mathcal{T}_b^{\text{train}} := [\![L_b^{\text{test}} - h, L_b^{\text{test}} - 1]\!], \tag{2.22}$$

and evaluate them with the b-th test set :

$$\mathcal{T}_b^{\text{test}} := [\![L_b^{\text{test}}, R_b^{\text{test}}]\!], \tag{2.23}$$

where $L_b^{\text{test}} := L^{\text{test}} + b \times s$ and $R_b^{\text{test}} := L_b^{\text{test}} + s - 1$.

Example 2. *For instance, consider that we want to evaluate the performances of a model in* 2016*, which corresponds to* $[\![L^{test}, R^{test}]\!]$ *and we choose, first to use* $h = 3$ *years in the training datasets, secondly to update the model every* $s = 4$ *weeks. The first test subset starts at* 0 *a.m. on January* 1^{st}*, ends at* 11 *p.m. on January 28, 2016 and is used to evaluate a model trained with data from January 2013 to December 2015. The second model is trained with data from February 2013 to January* 28^{th}*, 2016 and is evaluated with the data from* 0 *p.m. on January* 29^{th}*, 2016 to* 11 *p.m. on February* 25^{th}*, and so on until December 2016.*

In the end, we compute for each time period $\mathcal{T}_b^{\text{test}}$ with $b \in [\![0, B-1]\!]$ the quantities $\texttt{Mr}^2{}_b$, $\texttt{MMAPE}_b$, $\texttt{RMNMSE}_b$, $\texttt{WRNMSE}_b$ and $\texttt{MMAPE}_b$ defined in Section 2.7.2 for the forecasts $(\hat{\ell}_i^k)_{i=L_b^{\text{test}},\dots,R_b^{\text{test}},k=1,\dots,K}$ of the time series $(\ell_i^k)_{i=L_b^{\text{test}},\dots,R_b^{\text{test}},k=1,\dots,K}$. We also define the average performances over the batches :

$$\texttt{MMr}^2 = \frac{1}{B}\sum_{b=1}^{B} \texttt{Mr}^2{}_b, \tag{2.24}$$

$$\texttt{MMMAPE} = \frac{1}{B}\sum_{b=1}^{B} \texttt{MMAPE}_b, \tag{2.25}$$

$$\texttt{MRMNMSE} = \frac{1}{B}\sum_{b=1}^{B} \texttt{RMNMSE}_b. \tag{2.26}$$

Finally, to compare the predictions in different areas in a same aggregation level, we define the Mean RNMSE (**MRNMSE**) for the forecasts of one time series indexed with k :

$$\texttt{MRNMSE}_k = \frac{1}{B}\sum_{b=1}^{B} \texttt{RNMSE}_{k,b}, \tag{2.27}$$

where $\texttt{RNMSE}_{k,b}$ is the RNMSE defined in Equation (2.14) for time series k over the time period $\mathcal{T}_b^{\text{test}}$.

Size of the batches Two comments are in order after the presentation of the performance measures in Section 2.7.3. First, having a batch of results for repeated experiments is more convenient to compare results since the mean over the batches for different models might be quite close and having repeated samples lets use a Wilcoxon signed-rank test [Wilcoxon, 1992] to determine whether the differences are significant. This is why we do not compute the prediction errors over the entire test set at once.

Secondly, we had to determine the size of the batches. Although comparing the errors of different models for every single observation instant would generate more samples for the Wilcoxon test, these samples would be highly correlated, because the errors of one model at subsequent instants are generally highly correlated. Therefore, we had to determine a size of batch to aggregate the errors. For convenience, we chose it so that it matches the frequency s of the updates of the models.

2.8 Available equipment and ambitions

In this manuscript, we study load forecasting models able to provide daily at 23:59, forecasts for the next 24 hours of the hourly electricity demands in France. We describe in this section, the equipment used for our experiments.

Time constraints For all the models that we consider, the computational time necessary to make forecasts for 24 hours is never a problem. What matters is the time required to learn the models, and given the requirements of the TSO, this time should not exceed a few hours. For information, the existing national load forecasting models are estimated in a couple of minutes at most. In the end at the local level, our models required between 1 and 6 hours, depending on the exact choice of hyperparameters.

Computing power constraints Most experiments were performed on a High Performance Computing (**HPC**) cluster with 20 nodes (192 Go RAM each) and a total of about 300 CPU (3 GHz) but can be and were from time to time performed on a modern laptop (3.1 GHz, 16 Go RAM). In addition, 4 GPU were available and used to train and test neural networks or to occasionally leverage the simplicity of the *autograd* package [Maclaurin et al., 2015].

Memory constraints Although no memory constraint was imposed at the beginning of this work, experiments were never performed on a machine with more than 192 Go RAM. Experiments using around 60 Go RAM existed but were not suitable to large search of hyperparameters. Consequently, for convenience, experiments only occasionally exceeded 16 Go RAM.

2.9 Benchmarks

In order to compare our results with existing models, we consider three possibilities : the operational model used by the TSO, general purpose Machine Learning models and, recent works especially dedicated to load forecasting. We introduce all the benchmarks in this section.

2.9.1 Operational models

Aggregated load EDF developed during the 80s, before it was split into different entities in 2004 as described in Section 1.1.1, a model to forecast the national load [Bruhns et al., 2005; RTE, 2014]. The forecasts are published online [RTE, 2019a]. Experts consider that research papers in the electricity domain reached better performances only recently and this model is still in operation in major companies : its average MAPE for the day-ahead national load forecasting problem is 1.5 % [Antoniadis et al., 2012, Section 3.3].

However, this model has a large number of hyperparameters and its fine tuning required a lot of expert knowledge. An adaptation of this model to forecast the load of the administrative regions was proposed a few years ago but it is considered not to be suitable for local load predictions.

Although the forecasts of this historical model are made available on the website Eco2Mix [RTE, 2019a], we did not have access to this tool and cannot use it as a comparison for the local models that we are interested in. Besides, the exact conditions in which this model is used are slightly different from the ones that we consider. First, the forecasts are done at 17:00 instead of 23:59. Secondly, the temperatures in operational conditions are forecasts and not observations. Finally, the forecasts of the historical model are manually corrected *a posteriori* by forecasters and this procedure is unknown to us.

Local prediction Currently, the operational forecasts at the local levels are made by the TSO by distributing the regional forecasts to the sublevels based on historical proportions. This *top-down* method has been studied by Gross and Sohl [1990].

It seems however, based on expert knowledge, that modern Machine Learning tools perform better than the *top-down* method. Besides, we could not access the

precise model used by the TSO. As a consequence, we do not use the operational models as benchmarks.

2.9.2 Tree-based models

Tree-based models are an easy and remarkably fast forecasting tool, despite their difficult interpretation.

Random Forests In particular, Random Forests [Breiman, 2001] are widely used in the industry, were considered by Dudek [2015] for electricity load forecasting and allow to obtain quickly reasonable forecasts with :

- the instantaneous temperatures or cloud covers at different weather stations,
- the same weather conditions with a given delay (*e.g.* $\delta = 24$ or 48 hours),
- maximum and minimum of meteorological conditions over different time windows.
- the hour of the day,
- the day of the week,
- the day of the year,
- indicators of holidays, any annual or punctual event,
- the position of the sun (for the natural light),

We have not performed an exhaustive optimization of the hyperparameters of the random forests. However, the first results obtained showed that the high non-linearity in these models can help obtain reasonable forecasts.

XGBoost The algorithm XGBoost [Chen and Guestrin, 2016; Friedman, 2001], which is implemented in the XGBoost package [Chen et al., 2015], has also been competitive in data challenges so we considered it as a benchmark. It seems to benefit from the boosting since it performed most of the time better than Random Forests in our experiments. We also include it among the benchmarks in the first experiments of Section 3.5.2.

Remark 3. *Although there are multitask versions of these tree-based algorithms [Dumont et al., 2009], we observed that single-task versions applied to each time series independently perform better on the considered load forecasting problems. Note however that we have not searched exhaustively for the optimal hyperparameters.*

To summarize, tree-based models provide fast and flexible tools but the obtained models are extremely noisy and few generalization guarantees exist. Understanding and interpreting these highly non-linear models in a high-dimensional setting is intricate and a different topic.

Remark 4. *While in most models of this manuscript we include among the inputs the hour of the week, ranging from 0 to 167, tree-based models perform better with both the hour of the day and the day of the week.*

2.9.3 Neural networks

The advent of Neural Networks [Haykin, 1994, and references therein] led us to wonder what performances could be obtained for our problem with a short implementation that relies on the recent and remarkably powerful Deep Learning libraries. Besides, they were specifically studied for electricity load forecasting by Hippert et al. [2001]; Khotanzad et al. [1997]; Kiartzis et al. [1995]; Park et al. [1991].

We limited our experiments with neural networks by varying only a few hyperparameters and considering constant width network. We allowed the following hyperparameters to vary :

- the type of the layers : Fully Convolutional (FC) or ResNet Blocks,
- the depth of the network, that is the number of layers,
- the width of each layer, that is the number of units in a layer,
- the length of the training set,
- the weight decay parameters *i.e.* the regularization coefficients.

With the national load forecasting problems, we could quickly conclude that the networks benefited from longer training sets, which is a potential issue in a non-stationary context. Besides, the high dimension of the hyperparameters space and the slow optimization of the models made difficult the use the neural networks as benchmarks.

Since the first experiments, the performances of the fitted neural networks improved a bit for the national problem by exploring the space of hyperparameters. Although they are not yet competitive with tree-based models, they would certainly deserve a longer study.

In one sentence, neural networks were given important computing resources and did not outperform the other benchmarks, yet we have not concluded that this tool can be discarded for the load forecasting problem. Given their computational cost, they are not a convenient benchmark so far.

2.9.4 Related work - Generalized additive models

Generalized Additive Models (**GAM**) are an extension of the Generalized Linear Models (**GLM**) allowing the relationships between the inputs and the target variable to be non-linear. They have been studied specifically for the electricity load forecasting problem by the research team of EDF, at the national level by Pierrot and Goude [2011] and at the level of substations by Goude et al. [2013]. They are considered to provide state-of-the-art forecasts and indeed perform better with our database than the other benchmarks. We consequently describe them more precisely in this section where we restrict the scope to single-task models, multi-task models being discussed in Section 4.2.

GLM Generalized Linear Models (**GLM**) are a generalization proposed by Nelder and Wedderburn [1972] of linear regression to unify various statistical models [McCullagh and Nelder, 1983]. In a GLM, the distribution of the target variable $\mathsf{y} \in \mathbb{R}$ generalizes both the exponential families of distributions [Andersen, 1970; Darmois, 1935; Koopman, 1936; Pitman, 1936] that are defined with respect to a reference measure $\mu(d\mathsf{y})$ and have the form :

$$p(\mathsf{y}|\boldsymbol{\theta}) = k(\mathsf{y}) \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathsf{y}) - A(\boldsymbol{\theta})\right], \tag{2.28}$$

and the exponential dispersion models [Jørgensen, 1987] :

$$p(\mathsf{y}|\theta, \lambda) = h(\mathsf{y}, \lambda) \exp\left(\lambda \left[\theta \mathsf{y} - B(\theta)\right]\right), \tag{2.29}$$

where $s \in \mathbb{N}^*$, $\boldsymbol{\theta} \in \mathbb{R}^s$ and $\theta \in \mathbb{R}$ are *location parameters*, $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathbb{R}^s$ is the *canonical parameter*, the real-valued functions k and h are the *ancillary statistics*, $\boldsymbol{\phi}(\mathsf{y}) \in \mathbb{R}^s$ is the *sufficient statistic*, $A(\boldsymbol{\theta})$ and $B(\theta)$ are the real-valued *log-partition functions*, and $\lambda \in \mathbb{R}$ is a *scale parameter*.

More precisely, the target variable $\mathsf{y} \in \mathbb{R}$ in a GLM is assumed to follow a distribution in an overdispersed exponential family [Gelfand and Dalal, 1990] that generalizes Equation (2.28) and Equation (2.29) and whose density is given with respect to a reference measure $\mu(d\mathsf{y})$ by :

$$p(\mathsf{y}|\boldsymbol{\theta}, \tau) = h(\mathsf{y}, \tau) \exp\left(\frac{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathsf{y}) - A(\boldsymbol{\theta})}{\delta(\tau)}\right), \tag{2.30}$$

where $\tau \in \mathbb{R}$ is a *dispersion parameter* and δ is a real-valued *dispersion function*.

Furthermore, the mean value $\mathbb{E}[\mathsf{y}]$ of the target variable in a GLM depends on the input variables $(\xi_1, \ldots, \xi_D)$ through the link function $g : \mathbb{R} \to \mathbb{R}$ and coefficients $\beta_1, \ldots, \beta_D \in \mathbb{R}$ via the relationship :

$$\mathbb{E}[\mathsf{y}] = g^{-1}\left(\sum_{d=1}^{D} \xi_d \beta_d\right). \tag{2.31}$$

The Linear Regression corresponds to the special case where g is the identity function and the Ordinary Least Squares make the additional assumption that y follows a Gaussian distribution.

GAM In a GAM, the target variable y is related to the inputs with the more general structure :

$$\mathbb{E}[\mathsf{y}] = g^{-1}\left(\sum_{d=1}^{D} f_d(\xi_d)\right), \tag{2.32}$$

where for all $d \in [\![1, D]\!]$, the function $f_d : \mathbb{R} \to \mathbb{R}$ is an unspecified and possibly non-parametric function.

In addition to the flexibility provided by this structure, the GAM are motivated by the Kolmogorov-Arnold representation theorem [Arnold, 1957; Kolmogorov, 1957] which states that any continuous function Φ of the inputs $\xi_1, \ldots, \xi_D$ can be written as a finite composition of univariate functions :

$$\Phi(\xi_1, \ldots, \xi_D) = \sum_{e=0}^{E} g_e^{-1}\left(\sum_{d=1}^{D} f_{d,e}(\xi_d)\right) \tag{2.33}$$

with $E = 2D$. The GAM structure in Equation (2.32) corresponds to the restriction $E = 0$ *i.e.* with the outer sum dropped in Equation (2.33).

While the GAM contain a broad class of functions and general frameworks were studied for instance in [Breiman and Friedman, 1985; Friedman and Stuetzle, 1981], modeling assumptions generally require that the link function g is the identity or at least is known. Thereby, the relationship between the inputs and the target variable is additionally restricted in the models that we consider to a form that still generalizes linear regression :

$$\mathbb{E}[\mathsf{y}] = \sum_{d=1}^{D} f_d(\xi_d). \tag{2.34}$$

Estimation Several methods have been proposed to fit these models, see for instance a review in [Hastie and Tibshirani, 1990].

Tibshirani and Hastie [1987] and Hastie and Tibshirani [1990] study the estimation of the non-linear functions with the *back-fitting* algorithm, which is essentially a cyclic block coordinate descent algorithm, to iteratively increase the local likelihood. Instead, Wood [2017] developed the Penalized Iterative Re-Weighted Least Squares (P-IRLS) method, following the study of IRLS by Green [1984].

Implementation Penalized regression is implemented for *R* in the *MGCV* library [Wood and Wood, 2015] and was selected by Goude et al. [2013]; Pierrot and Goude [2011]; Wood et al. [2015] for electricity forecasting to minimize the following objective :

$$\mathbb{E}\left[\left(\mathsf{y} - \sum_{d=1}^{D} f_d(\xi_d)\right)^2\right] + \sum_{d=1}^{D} \lambda_d \int \|f_d''\|^2, \tag{2.35}$$

with the methodology described in [Wood, 2004, 2011], and where the functions $(f_d)_{d=1,\dots,D}$ are parametrized with linear combinations of pre-determined basis functions, often piecewise polynomials with a given number of knots. Note that by setting for all $d \in [\![1, D]\!]$, $\lambda_d \to +\infty$, the penalization of the curvature in Equation (2.35) constrain the univariate functions $(f_d)_{d=1,\dots,D}$ to be linear and the model becomes a GLM.

An interesting feature of the *MGCV* library is the automatic selection of the regularization coefficients λ_d, based on a Generalized Cross-Validation Criteria (**GCV**) [Craven and Wahba, 1978].

Extension to multivariate functions Locally, the estimation of the functions $(f_d)_{d=1,\dots,D}$ relies on the density of the data points given in the training set and it is well-known that smoothing techniques break down in high dimensions if the functions $(f_d)_{d=1,\dots,D}$ are multivariate, *i.e.* have vector as arguments [Friedman and Stuetzle, 1982]. For this purpose, Wahba [1980] proposed the thin-plate splines to model response surfaces.

Load forecasting GAM Both for the national and local forecasting problems, Pierrot and Goude [2011] and Goude et al. [2013] estimate several GAM for different

times of the day. At the national level, Pierrot and Goude [2011] build 24 models, one for each hour, while at the local level Goude et al. [2013] estimate 144×2000 models since they considered loads measured every 10 minutes at about 2000 substations. In addition to the inputs considered for the tree-based models in Section 2.9.2, they include the timestamp `t` which is the number of seconds since a chosen instant in time. Pierrot and Goude [2011] also include interactions between pairs of inputs with thin plate-splines in the national model, that are implemented in the *MGCV* library.

Impossible formal comparison Data challenges have been proposed with the objective of comparing different forecasting models based on North American data [Hong and Fan, 2016; Hong et al., 2014]. However, the exact problem that they consider is different from day-ahead load forecasting and the predictions of each group of participants are unknown, like the locations of the substations and the weather stations that correspond to a higher level of aggregation than the French substations.

As for French electricity datasets, most of them are confidential : there is no large public dataset to formally compare models. We do not have access to the list of substations considered by Goude et al. [2013] for the local load forecasting problem, or to the selected special tariffs information. Besides, the time period in our dataset is different from theirs.

We consequently considered their modeling and for reproducibility, we give in Appendix D the exact formulas that we used in *R* with the *MGCV* library [Wood and Wood, 2015], to obtain benchmarks with GAM.

Chapter 3

Independent models

We only consider in this chapter load forecasting models where each time series is predicted independently from the others. After introducing a mathematical framework for a general form single-task discriminative model, we propose in Section 3.4 middle and short-term instances to compare with the benchmarks introduced in Section 2.9. Experiments and their analyses are presented in Section 3.5 and Section 3.6.

3.1 Feature engineering

As seen in Section 2.3, an appropriate modelling of the load requires nonlinearity. Consequently, we propose feature transformations with sets of basis functions adapted to the four categories of inputs distinguished in Table 3.1 : the indicators, the timestamp, the bounded acyclic inputs (past loads and weather conditions) and the cyclic inputs (hour and day of the year).

Origin	Inputs	Categories
Calendar	Hour of the week Day of the year	Cyclic
	Sun is up 11 French holidays Christmas period	Indicators
	Timestamp	Timestamp
Weather	Temperatures Cloud cover	Acyclic
Endogenous	Past loads*	

* Only for the short-term models

TABLE 3.1: Classification of the inputs
The timestamp has its own category. It is indeed conceptually different from the other variables because unlike the temperatures or the loads whose whole range of possible values is encountered in the training sets, the forecasts are an extrapolation with respect to the timestamp.

From the results of the tree-based models of Section 2.9.2 and the bivariate conditional expectations presented in Section 2.3.3, we expect the model to require

at least one-dimensional effect and interactions between pairs of inputs. So, we introduce both univariate and bivariate features to eventually define a standard linear model.

For clarity, the original data time series are called the inputs. These inputs are transformed with functions called features and the instances of these features are the covariates. A summary of the whole procedure leading to the minimization Problem (3.25) is summarized in Appendix C.

Indicators and timestamp There is no transformation of the binary variables in our models. Still, given an input $\xi \in \{0, 1\}$, we write $\phi_0(\xi) := \xi$ the corresponding feature to fit the general framework.

Besides, the only transformations of the timestamp $\mathsf{t} \in \mathbb{R}$ that we consider are polynomial, to model and extrapolate a long-term trend. For polynomials of degree at most $p_{\text{timestamp}} \in \mathbb{N}^*$, we write the corresponding vector of features $\boldsymbol{\phi}_{\text{timestamp}}(\mathsf{t}) := (\mathsf{t}, \ldots, \mathsf{t}^{p_{\text{timestamp}}})$, typically we set $p_{\text{timestamp}} = 1$ or $p_{\text{timestamp}} = 2$.

For all the other inputs, we create features with univariate splines.

3.1.1 Univariate splines

Basis functions An ideal family of basis functions for a nonlinear transformation of the inputs contains elements that simultaneously :

- are regular,
- have a simple analytical form,
- are orthogonal for the scalar product related to the distribution of the inputs,
- have a localized support.

Different families were proposed for modeling but none satisfies all the sought conditions. We compromise and use the framework provided by the cardinal B-splines [De Boor et al., 1978; Eilers and Marx, 1996]. They satisfy 3 out of the 4 criteria because their supports are not disjoint and the orthogonality condition is not satisfied.

The splines and their different variants are particularly adequate with well-studied approximation properties [Schumaker, 2007; Wahba, 1990]. Like a vast majority of the basis function encountered in the literature, they are defined by a set of knots. With splines, the modeled function is allowed to have discontinuous derivatives at these knots.

Different approaches were proposed to select the knots. They can be fixed and uniformly distributed or computed from the quantiles of the data. Alternatively, Friedman et al. [1991] proposed a forward selection algorithm for Multivariate Adaptive Regression Splines (**MARS**). Adaptive models have alternatively been considered with other iterative procedures [Zhou and Shen, 2001], with the Trend Filtering models (**TF**) [Tibshirani et al., 2014] and with the Locally Adaptive Regression Splines (**LARS**) [Mammen et al., 1997] that consider combinations of the elements of the truncated power basis functions and lead to piecewise polynomials with pieces as large as possible. Because the supports of the truncated power basis functions

are not compact, Bakin et al. [1999] later adapted the MARS model to B-splines in the **BMARS** framework, supposedly leading to better conditioned design matrices.

Cardinal B-splines The B-splines are piecewise polynomial with possible discontinuities of the derivatives localized in a finite set of knots. They are additionally called cardinal if the knots are equidistant. Although their supports are not disjoint, this family is particularly suitable for approximation since any spline function can be written as a linear combination of B-splines.

More precisely, we build basis functions with the cardinal B-spline with degree 1 and support $[0, 2]$,

$$B^1 : \xi \mapsto (\xi)_+ - 2(\xi - 1)_+ + (\xi - 2)_+, \tag{3.1}$$

where $(\zeta)_+ := \max(\zeta, 0)$. There are more regular B-splines with higher degrees and larger supports, that we illustrate in Figure 3.1. For instance the cubic splines [Stone and Koo, 1985] insist on the regularity of the basis functions but have larger supports. The cardinal B-spline with degree δ and support $[0, \delta + 1]$ is given by :

$$B^\delta : \xi \mapsto \frac{1}{\delta!} \sum_{m=0}^{\delta+1} (-1)^m \binom{\delta+1}{m} (\xi - m)_+^\delta, \tag{3.2}$$

where $\binom{\delta+1}{m} := \frac{(\delta+1)!}{(\delta+1-m)!m!}$ is the binomial coefficient. For any $\delta \in \mathbb{N}$, the function B^δ is $\mathcal{C}^{\delta-1}$ and piecewise $\mathcal{C}^\delta$. Using piecewise linear splines with $\delta = 1$ instead of splines with a higher degree has the notable advantage of producing continuous and sparser representations.

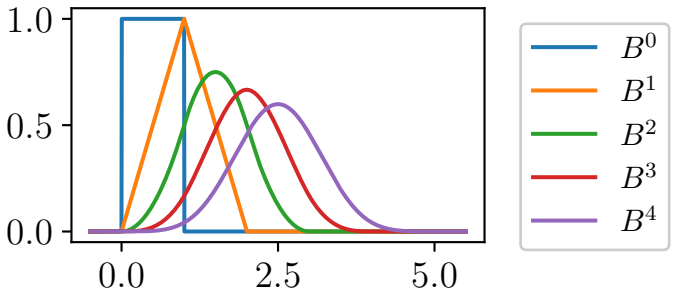


FIGURE 3.1: The cardinal B-splines B^δ for $\delta = 0, 1, 2, 3, 4$.

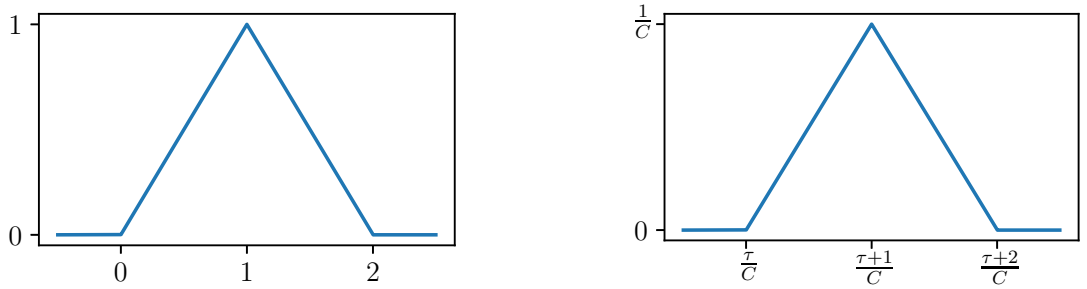
Family of basis splines We generate a family of basis functions thanks to compositions of affine functions with B^1. We follow the ideas of a multi-resolution approximation [Forster, 2011], with possibly non-dyadic cuts. This last point is relevant for hours of the day or hours the week since $24 = 2^3 \times 3$ and $168 = 7 \times 24$ are not powers of 2.

Consider a sequence of cuts $(c_r)_{r\in\mathbb{N}} \in (\mathbb{N}\backslash\{0, 1\})^{\mathbb{N}}$ and a level of detail $\ell \in \mathbb{N}\backslash\{0, 1\}$, we define the granularity $C := \prod_{r=1}^{\ell} c_r$. It is inversely proportional to the

support width of the splines that we build. Given a translation parameter $\tau \in \mathbb{Z}$, we define the perspective function $B_{\tau,C}$ with support $[\frac{\tau}{C}, \frac{\tau+2}{C}]$ as :

$$B_{\tau,C} : \xi \mapsto \frac{1}{C} B^1(C\xi - \tau). \tag{3.3}$$

As illustrated in Figure 3.2b, the support of $B_{\tau,C}$ is centered at $\frac{\tau+1}{C}$.



(a) The cardinal B-spline B^1. **(b)** The composed function $B_{\tau,C}$.

FIGURE 3.2: The transformed Cardinal B-spline B^1.

Restriction to $[0, 1]$ To describe a general procedure, we consider that the inputs have already been affinely transformed so that the cyclic inputs with original values in $[0, c]$, where c is the maximum value (*e.g.* the value of c is $167 = 7 \times 24 - 1$ for the hour of the week) now lie in $[0, 1 - \frac{1}{c+1}]$ and the other inputs have been transformed so that the minimum is 0 and the maximum is 1, this is detailed in Appendix C. Thus, we only select those elements whose support has a non-trivial intersection with the interval $[0, 1]$:

$$\mathcal{S}^C := \{B_{\tau,C},\ \tau \in [\![-1, C-1]\!]\}\,. \tag{3.4}$$

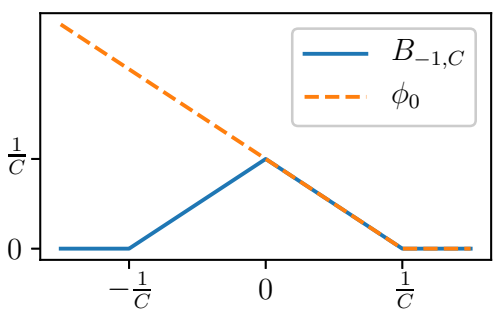
The family $\mathcal{S}^C$ spans the set of piecewise linear continuous functions that are zero outside $[-\frac{1}{C}, 1 + \frac{1}{C}]$ and whose derivative may be discontinuous at the knots $\{\frac{m}{C},\ m \in [-1, C+1]\}$. We adapt it for the acyclic and cyclic inputs classified in Table 3.1 : to anticipate extrapolation in the first case and to satisfy the additional constraint in the second case.

Acyclic features Generally, estimators near the boundaries of the observed domain tend to be erratic, which lead Friedman et al. [2001, Section 5.2.1] to consider the natural cubic splines that are piecewise third-order polynomials with the additional condition, from which the adjective *natural* is coined, that the second-order derivative is zero on the two edges of the domain, the extrapolation outside the observed domain being linear.

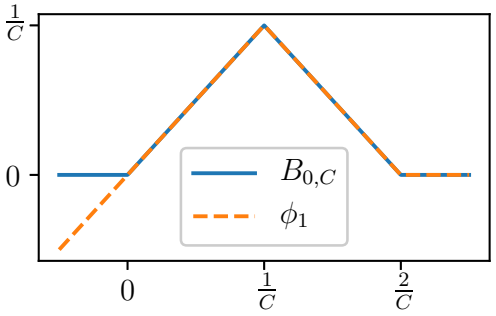
In our case, although the training data is affinely transformed to lie in the interval $[0, 1]$, the same transformations on new unseen data might have values outside $[0, 1]$. Following the ideas of the natural splines, we choose instead of $\mathcal{S}^C$, a family of functions whose span is the set of piecewise linear functions with possible discontinuities of the derivative in the set of knots $\{\frac{m}{C},\ m \in [1, C-1]\}$ and that are linear

outside of $[\frac{1}{C}, 1-\frac{1}{C}]$. Let $\mathcal{A}^C$ denote this family of $C+1$ continuous transformations for acyclic inputs, shown in Figure 3.3 :

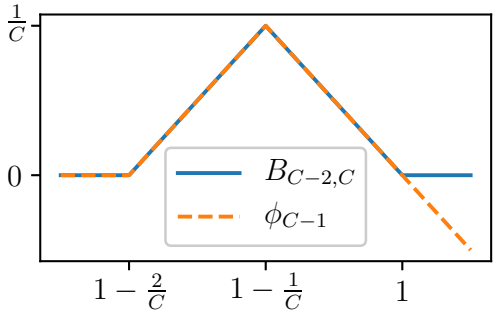
$$\begin{aligned}\mathcal{A}^C = &\{\ \phi_0 &&: \xi \mapsto \max(0, \frac{1}{C} - \xi)\ \}\\ &\cup \{\ \phi_1 &&: \xi \mapsto \min(B_{0,C}(\xi), \xi)\ \}\\ &\cup \{\ \phi_{\tau+1} &&: \xi \mapsto B_{\tau,C}(\xi),\ \tau \in [\![1, C-3]\!]\ \}\\ &\cup \{\ \phi_{C-1} &&: \xi \mapsto \min(B_{C-2,C}(\xi), 1-\xi)\ \}\\ &\cup \{\ \phi_C &&: \xi \mapsto \max(0, \xi - 1 + \frac{1}{C})\ \}\,.\end{aligned}$$



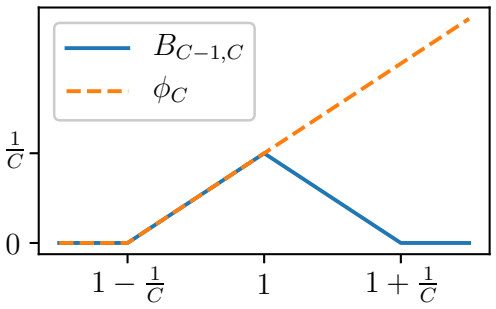
(a) $B_{-1,C} \in \mathcal{S}^C$ and $\phi_0 \in \mathcal{A}^C$



(b) $B_{0,C} \in \mathcal{S}^C$ and $\phi_1 \in \mathcal{A}^C$



(c) $B_{C-2,C} \in \mathcal{S}^C$ and $\phi_{C-1} \in \mathcal{A}^C$



(d) $B_{C-1,C} \in \mathcal{S}^C$ and $\phi_C \in \mathcal{A}^C$

FIGURE 3.3: Family of univariate acyclic splines
Modification of the acyclic basis functions for extrapolation purposes.

Finally, we denote by $\boldsymbol{\phi}$ the concatenation of the linearly independent elements $(\phi_0, \ldots, \phi_C)$ of $\mathcal{A}^C$. Since $\sum_{j=0}^{C} \phi_j$ equals the constant function $\xi \mapsto \frac{1}{C}$, the union of such families for different inputs will not be linearly independent. It is the case for instance if we consider an additive model with at least 2 temperatures as inputs, each one being associated to a different vector of features.

Cyclic features For a cyclic input, there is an additional constraint but extrapolation is not a concern anymore. Among the functions of $\mathcal{S}^C$, only $B_{-1,C}$ and $B_{C-1,C}$ do not have a trivial cyclic extension. However, we see in in Figure 3.4 that they are naturally replaced by merging them.

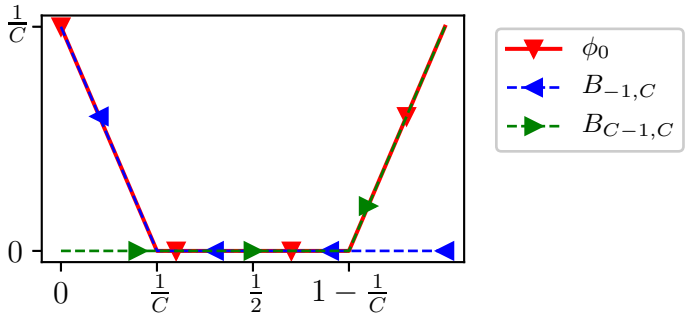


FIGURE 3.4: Family of univariate cyclic splines
Modification of the basis to satisfy the cyclic constraint. The pair $(B_{-1,C},\, B_{C-1,C})$ in $\mathcal{S}^C$ is substituted with ϕ_0 in $\mathcal{C}^C$.

Therefore, we define the family of cyclic basis functions :

$$\begin{aligned}\mathcal{C}^C = \quad & \{\, \phi_0 \quad : \xi \in \mathbb{R}/\mathbb{Z} \mapsto \max[B_{-1,C}(\xi),\, B_{C-1,C}(\xi)] \,\} \\ & \cup \{\, \phi_{\tau+1} : \xi \in \mathbb{R}/\mathbb{Z} \mapsto B_{\tau,C}(\xi),\ \tau \in [\![0, C-2]\!] \,\}\,.\end{aligned}$$

Because of the additional constraint, the number of elements in $\mathcal{C}^C$ is only C. We denote by $\boldsymbol{\phi}$ the multivariate feature obtained by concatenating the elements $(\phi_0, \ldots, \phi_{C-1})$ of $\mathcal{C}^C$. Note that for an input with discrete values in $[0, \frac{1}{c+1}, \ldots, 1 - \frac{1}{c+1}]$ where $c \in \mathbb{N}$, we can build indicators from the representations above based on splines if we choose a sufficient level of detail $C = c + 1$. This will be of particular interest when considering as input the hour of the week $\mathsf{h} \in [\![0, 167]\!]$ with $c = 167$: having an indicator for each hour of the week means that the set of basis functions spans all functions of these discrete values.

3.1.2 Interactions

Bivariate features To allow interactions in the model between the different inputs, we build bivariate features with tensor products of univariate features [Bakin et al., 1999; Binev et al., 2007]. Consider two inputs $\xi, \zeta \in [0, 1]$, for instance the past load and the hour of the week, and the associated vector of features $\boldsymbol{\phi} \in (\mathbb{R}^{\mathbb{R}})^p$ and $\boldsymbol{\psi} \in (\mathbb{R}^{\mathbb{R}})^q$ built in Section 3.1.1, where $p, q \in \mathbb{N}^*$. We define the interaction features with the tensor product :

$$\boldsymbol{\phi} \otimes \boldsymbol{\psi} \in (\mathbb{R}^{\mathbb{R}})^{p,q}. \tag{3.5}$$

Given two inputs $\xi, \zeta \in \mathbb{R}$, it is convenient to see the covariates associated to this interaction as a matrix :

$$\boldsymbol{\Phi}(\xi, \zeta) := \boldsymbol{\phi}(\xi)\boldsymbol{\psi}(\zeta)^T \in \mathbb{R}^{p,q}. \tag{3.6}$$

Thus, any linear combination of these covariates with a coefficient matrix $\boldsymbol{M} \in \mathbb{R}^{p,q}$ can be written :

$$\langle \boldsymbol{\Phi}(\xi, \zeta), \boldsymbol{M} \rangle. \tag{3.7}$$

3.2 Additive model

With the univariate and bivariate features, we build a general-form additive model to apply to a vector of inputs $\boldsymbol{\xi} := (\xi_1, \dots, \xi_D) \in \mathbb{R}^D$.

First, consider a subset of the inputs $\mathcal{U} \subset [\![1, D]\!]$ for the univariate features and $d \in \mathcal{U}$. After setting a granularity $C_d \in \mathbb{N}^*$ for the input d, we write the corresponding vector of features $\boldsymbol{\phi}_d \in (\mathbb{R}^{\mathbb{R}})^{p_d}$ where $p_d = C_d + 1$ if d corresponds to an acyclic input, $p_d = C_d$ if d corresponds to a cyclic input, $p_d = 1$ if d is an indicator and $p_d = p_{\text{timestamp}}$ if d corresponds to the timestamp. We denote the total number of univariate features

$$p_{\mathcal{U}} := \sum_{d \in \mathcal{U}} p_d. \tag{3.8}$$

Given a subset of interactions $\mathcal{B} \subset [1,\ D]^2$ and $(d, e) \in \mathcal{B}$, we write p_d and p_e the size of the associated univariate features used to build the corresponding matrix of features as in Section 3.1.2 with a tensor product :

$$\boldsymbol{\Phi}_{d,e} := \boldsymbol{\phi}_d \otimes \boldsymbol{\phi}_e \in (\mathbb{R}^{\mathbb{R}})^{p_d, p_e}, \tag{3.9}$$

so that the total number of bivariate features is :

$$p_{\mathcal{B}} = \sum_{(d,e) \in \mathcal{B}} p_d p_e. \tag{3.10}$$

We define the additive model $\mathcal{M}_{\mathcal{U},\mathcal{B}}$ with :

$$p := 1 + p_{\mathcal{U}} + p_{\mathcal{B}} \tag{3.11}$$

coefficients as the following set of hypotheses :

$$\begin{aligned} (\xi_1, \dots, \xi_D) \mapsto \quad & \beta_0 + \sum_{d \in \mathcal{U}} \langle \boldsymbol{\phi}_d(\xi_d), \boldsymbol{\beta}_d \rangle + \sum_{(d,e) \in \mathcal{B}} \langle \boldsymbol{\Phi}_{d,e}(\xi_d, \xi_e), \boldsymbol{M}_{d,e} \rangle \\ \text{s.t.} \quad & \beta_0 \in \mathbb{R}, \quad \forall\ d \in \mathcal{U},\ \boldsymbol{\beta}_d \in \mathbb{R}^{p_d}, \quad \forall (d,e) \in \mathcal{B},\ \boldsymbol{M}_{d,e} \in \mathbb{R}^{p_d, p_e}. \end{aligned} \tag{3.12}$$

Equivalently, denoting the univariate effects :

$$f_d : \xi_d \mapsto \langle \boldsymbol{\phi}_d(\xi_d), \boldsymbol{\beta}_d \rangle, \tag{3.13}$$

where $(f_d)_{d \in \mathcal{U}}$ are piecewise linear functions, and the bivariate effects :

$$g_{d,e} : (\xi_d, \xi_e) \mapsto \langle \boldsymbol{\Phi}_{d,e}(\xi_d, \xi_e), \boldsymbol{M}_{d,e} \rangle, \tag{3.14}$$

where $(g_{d,e})_{(d,e) \in \mathcal{B}}$ are defined with linear combinations of tensor products, the model $\mathcal{M}_{\mathcal{U},\mathcal{B}}$ has the concise form

$$(\xi_1, \dots, \xi_{D)} \mapsto \quad \beta_0 + \sum_{d \in \mathcal{U}} f_d(\xi_d) + \sum_{(d,e) \in \mathcal{B}} g_{d,e}(\xi_d, \xi_e). \tag{3.15}$$

3.3 Formulation of the optimization problem

To learn the coefficients in the model $\mathcal{M}_{\mathcal{U},\mathcal{B}}$, we classically write the minimization of the mean squared error on a training set as an optimization problem. Let p denote the number of covariates given in Equation 3.11 and $\mathbf{x} \in \mathbb{R}^p$ be the random vector obtained by concatenating all the covariates built with the features from the inputs $(\xi_1, \ldots, \xi_D) \in \mathbb{R}^D$:

$$\mathbf{x} := \left(1, [\boldsymbol{\phi}_d(\xi_d)]_{d\in\mathcal{U}}, [\boldsymbol{\Phi}_{d,e}(\xi_d, \xi_e)]_{(d,e)\in\mathcal{B}}\right). \tag{3.16}$$

Similarly, let $\boldsymbol{b} \in \mathbb{R}^p$ denote the coefficient vector obtained by concatenating the coefficients assigned to each group of features :

$$\boldsymbol{b} := \left(\beta_0, [\boldsymbol{\beta}_d]_{d\in\mathcal{U}}, [\boldsymbol{M}_{d,e}]_{(d,e)\in\mathcal{B}}\right). \tag{3.17}$$

The model (3.12) can be decomposed in two steps : the feature engineering of Section 3.1 and the linear combination of the covariates in Section 3.2, that can be written as :

$$\mathbf{x} \in \mathbb{R}^p \mapsto \mathbf{x}^T\boldsymbol{b} \quad \text{s.t.} \quad \boldsymbol{b} \in \mathbb{R}^p. \tag{3.18}$$

The data fitting term Consider a training set of observation instants $i \in [\![1, n]\!]$ with $n \in \mathbb{N}^*$ and a target time series $(y_i)_{i=1,\ldots,n} \in \mathbb{R}^n$. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ denote the covariates for instant i. Minimizing the squared error on the training set of the model $\mathcal{M}_{\mathcal{U},\mathcal{B}}$ amounts to minimizing with respect to $\boldsymbol{b} \in \mathbb{R}^p$ the data fitting term :

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{b})^2. \tag{3.19}$$

The regularization In order to avoid overfitting, we consider adding a regularization $\omega : \mathbb{R}^p \to \mathbb{R}_+$ and write the optimization problem :

$$\min_{\boldsymbol{b}\in\mathbb{R}^p} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{b})^2 + \omega(\boldsymbol{b}). \tag{3.20}$$

Since the coefficient vector $\boldsymbol{b} \in \mathbb{R}^p$ is the concatenation of coefficient vectors assigned to different groups of features, we use regularizations with the additive form :

$$\omega : [\beta_0, (\boldsymbol{\beta}_d)_{d\in\mathcal{U}}, (\boldsymbol{M}_{d,e})_{(d,e)\in\mathcal{B}}] \mapsto \quad \lambda_0\gamma_0(\beta_0) + \sum_{d\in\mathcal{U}}\lambda_d\gamma_d(\boldsymbol{\beta}_d) + \sum_{(d,e)\in\mathcal{B}}\lambda_{d,e}\gamma_{d,e}(\boldsymbol{M}_{d,e}),$$

where $\lambda_0 \geq 0$, for all $d \in \mathcal{U}$, $\lambda_d \geq 0$ and for all $(d, e) \in \mathcal{B}$, $\lambda_{d,e} \geq 0$.

Structured coefficients Adding the interactions between univariate features may significantly increase the statistical complexity of the model and exposes it to overfitting. To counteract this effect, we will add, given a matrix $\boldsymbol{M} \in \mathbb{R}^{p,q}$ associated to an interaction between two inputs ξ and ζ, three possible structural constraints. Thus, we distinguish :

- the unstructured case without any structural constraint. Therefore, the matrix $\boldsymbol{M}$ has $p \times q$ free parameters.

- the low-rank constraint :
$$\text{rank}(\boldsymbol{M}) \leq R, \tag{3.21}$$
with $R \in \mathbb{N}^*$, implying the existence of $\boldsymbol{U} \in \mathbb{R}^{p,R}$ and $\boldsymbol{V} \in \mathbb{R}^{q,R}$ such that $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^T$. With the notations of Section 3.1.2, this implies that the interaction can be written :
$$\langle \boldsymbol{\phi}(\xi)\boldsymbol{\psi}(\zeta)^T, \boldsymbol{M} \rangle = \sum_{r=1}^{R} [\boldsymbol{\phi}(\xi)^T \boldsymbol{u}^{(r)}][\boldsymbol{\psi}(\zeta)^T \boldsymbol{v}^{(r)}], \tag{3.22}$$
where $\boldsymbol{u}^{(r)}$ and $\boldsymbol{v}^{(r)}$ are the r-th columns of respectively $\boldsymbol{U}$ and $\boldsymbol{V}$. We use this form in our implementation and this leads to $(p+q)R$ parameters associated to this interaction. In this case, we use a penalization of the form $(\boldsymbol{U}, \boldsymbol{V}) \mapsto \omega_1(\boldsymbol{U}) + \omega_2(\boldsymbol{V})$ instead of penalizing the matrix $\boldsymbol{M}$.

- the sesquivariate constraint
$$\boldsymbol{M} = \boldsymbol{\beta}\boldsymbol{v}^T + \boldsymbol{u}\boldsymbol{\gamma}^T \tag{3.23}$$
where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient vector associated to the univariate feature of input ξ, $\boldsymbol{\gamma} \in \mathbb{R}^q$ is the coefficient vector associated to the univariate feature of input ζ and $\boldsymbol{v} \in \mathbb{R}^q$, $\boldsymbol{u} \in \mathbb{R}^p$. Thus, the sum of the univariate effects $\langle \boldsymbol{\phi}(\xi), \boldsymbol{\beta} \rangle + \langle \boldsymbol{\psi}(\zeta), \boldsymbol{\gamma} \rangle$ and this bivariate effect $\langle \boldsymbol{\phi}(\xi)\boldsymbol{\psi}(\zeta)^T, \boldsymbol{\beta}\boldsymbol{v}^T + \boldsymbol{u}\boldsymbol{\gamma}^T \rangle$ has the general form :
$$(\xi, \zeta) \mapsto F(\xi)[1 + g(\zeta)] + G(\zeta)[1 + f(\xi)]. \tag{3.24}$$
Note that for this constraint to be possible, the dimension must be the same for the univariate features and the bivariate features *i.e.* the matrix $\boldsymbol{M}$ must have the same number of rows as $\boldsymbol{\beta}$. With this constraint, the rank of $\boldsymbol{M}$ is at most 2 and the number of parameters associated to this interaction is then only $p + q$. In this case, we penalize the vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ instead of penalizing the matrix M.

The low-rank constraint (3.21) is introduced as an intermediary, if $R \geq 2$, between the unstructured case and the sesquivariate constraint (3.23) but empirically, it was never better than both, even for $R = 1$, so we do not use it.

Independent models The above formulation is immediately generalized to a multi-task setting where each task is learned independently from the others. Consider $K \in \mathbb{N}^*$ target series $(\boldsymbol{y}^{(k)})_{k=1,\dots,K}$ where for each $k \in [\![1, K]\!]$ we have $n \in \mathbb{N}^*$ observations : $\boldsymbol{y}^{(k)} \in \mathbb{R}^n$.

In the problems that we consider, some covariates are called *common* because they are used by all the tasks (*e.g.* the hour) and others are called *local* as only some tasks use them (*e.g.* the past loads or the temperatures at a given weather station). We denote their numbers respectively $p_0 \in \mathbb{N}$ and $p - p_0 \in \mathbb{N}$. Besides, we assume that all tasks have the same numbers of *local* covariates. We also denote

$\boldsymbol{X}^{(0)} \in \mathbb{R}^{n,p_0}$ the *common* design matrix and $\boldsymbol{Z}^{(k)} \in \mathbb{R}^{n,p-p_0}$ the *local* design matrix of each task $k \in [\![1, K]\!]$.

Similarly, for each task $k \in [\![1, K]\!]$, the coefficient vector associated to the *common* covariates is denoted $\boldsymbol{a}^{(k)} \in \mathbb{R}^{p_0}$ and the coefficient vector for the *local* covariates is denoted $\boldsymbol{c}^{(k)} \in \mathbb{R}^{p-p_0}$. We write the collective yet independent optimization problem as :

$$\min_{\boldsymbol{A}\in\mathbb{R}^{p_0,K},\boldsymbol{C}\in\mathbb{R}^{p-p_0,K}} \quad \frac{1}{2n}\sum_{k=1}^{K} \left\|\boldsymbol{y}^{(k)} - \boldsymbol{X}^{(0)}\boldsymbol{a}^{(k)} - \boldsymbol{Z}^{(k)}\boldsymbol{c}^{(k)}\right\|_2^2 + \omega_k(\boldsymbol{a}^{(k)}, \boldsymbol{c}^{(k)}). \tag{3.25}$$

For each $k \in [\![1, K]\!]$, we define $\boldsymbol{b}^{(k)} \in \mathbb{R}^p$ as the concatenation of $\boldsymbol{a}^{(k)}$ and $\boldsymbol{c}^{(k)}$:

$$\boldsymbol{b}^{(k)} := \begin{bmatrix} \boldsymbol{a}^{(k)} \\ \boldsymbol{c}^{(k)} \end{bmatrix}, \tag{3.26}$$

so that the column vector $\boldsymbol{b}^{(k)}$ is the analogous for task k of the vector $\boldsymbol{b}$ defined for the single-task setting in Equation (3.17) and $\boldsymbol{y}^{(k)} \in \mathbb{R}^n$ is the analogous of the time series $(y_i)_{i=1,\dots,n} \in \mathbb{R}^n$ defined for Equation (3.19).

3.4 Application to the load forecasting problem

Target variables In Section 2.5.2, we introduced 5 partitions of the $\mathcal{K}$ substations that correspond to different levels of aggregation. To instantiate a load forecasting model, consider one element $Z_k \subset [\![1, \mathcal{K}]\!]$ in these partitions, that corresponds to a subset of the substations, with cardinal $|Z_k|$. Like in Equation (2.6), we denote $(\mathsf{r}_\kappa)_{\kappa\in Z_k} \in \mathbb{R}^{|Z_k|}$ the loads of the substations in the area Z_k and we define the aggregated load of this area :

$$\ell_k := \sum_{\kappa\in Z_k} \mathsf{r}_\kappa \in \mathbb{R}. \tag{3.27}$$

In this section, we introduce a parametrization to model this aggregated load, that we denote for simplicity ℓ. Its estimator is denoted $\hat{\ell}$.

There is no reason for the parametrization to be the same at different aggregation levels and indeed, we found empirically that they should be different. Consequently, we restrict this section to the parametrization of the national models and the parametrization for the models at the level of substations are given in the three Tables F.2 - F.4. These choices are then justified in Sections 3.5 and 3.6.

Inputs The calendar inputs given to the middle-term model are the timestamp t, the hour of the week h and the day of the year d. We also use a unique binary indicator 1_{hld} for the 11 French holidays and two others 1_{hld^-} and 1_{hld^+} for the days respectively preceding and following a holiday. We denote 1_{xmas} an indicator of the two weeks around Christmas and New Year's Eve and 1_{sun} an indicator of the daytime, that is 1 between the sunset and the sunrise measured in Paris and 0 otherwise.

In Section 2.5.2, the subset of substation Z_k was associated with a subset of the weather stations that we denote $[\![1, S]\!]$ with $S \in \mathbb{N}^*$. For $s \in [\![1, S]\!]$ and $\delta \in \{24, 48\}$, we write T^s the corresponding temperature, $\mathsf{T}^s_{-\delta}$ the temperature with a delay of δ hours, $\overline{\mathsf{T}}^s_{-\delta}$ the maximum temperature over the δ last hours, and $\underline{\mathsf{T}}^s_{-\delta}$ the minimum. Finally, the instantaneous cloud cover is denoted c^s.

For short-term forecasts, the models are additionally aware of the past loads $\ell_{-\delta}$ with a delay of δ hours, for $\delta \in \{24, 48\}$. An enumeration of the inputs to the short-term model is given in Table 3.2.

	Category	Name	Symbol
Date and time	cyclic	hour of the week	h
		day of the year	d
	indicators	holidays	1_{hld}
		days before a holiday	1_{hld^-}
		days after a holiday	1_{hld^+}
		Christmas period	1_{xmas}
		Sun is up	1_{sun}
	absolute time	timestamp	t
Weather	acyclic	temperatures	T^s
		δ hours-delayed temperatures	$\mathsf{T}^s_{-\delta}$
		maximum over a δ hours window	$\overline{\mathsf{T}}^s_{-\delta}$
		minimum over a δ hours window	$\underline{\mathsf{T}}^s_{-\delta}$
		cloud covers	c^s
Past loads	acyclic	δ hours-delayed load	$\ell_{-\delta}$

TABLE 3.2: Inputs to the short-term load forecasting model
There are as many copies of the weather-related inputs as weather stations $s \in [\![1, S]\!]$ and one copy of the past information for each $\delta \in \{24, 48\}$.

Univariate features To build features as in Section 3.1, we decide of a granularity for each input. Based on preliminary experiments, we consider the hour of the week instead of considering separately the hour of the day and the day of the week like Goude et al. [2013]; Thouvenot [2015]. For this input, we choose 168 knots for the hour of the week because of the high frequencies of the expected load conditioned on this variable (*c.f.* Figure 2.10). This is equivalent with using an indicator for each hour of the week.

For the other continuous inputs, we expect smoother univariate effects given the empirical conditional expectations presented in Section 2.3. The granularity is set to 128 for the day of the year, to 16 for the temperature-related univariate features and, to 4 for the past loads in the short-term models, leading respectively to 128, 17 and 5 knots because the temperatures and the past loads are not cyclic. Finally, we add a linear function for the timestamp t.

The coefficients for the hour of the week h and the timestamp t are penalized with the Ridge regularization :

$$\|\cdot\|_F^2 : \boldsymbol{\theta} \in \mathbb{R}^p \mapsto \frac{1}{2} \sum_{i=1}^{p} \theta_i^2. \tag{3.28}$$

For the other coefficients, we use the smoothing spline regularization Ω_{S^2}, like in [Fan and Hyndman, 2011; Goude et al., 2013; Pierrot and Goude, 2011; Thouvenot, 2015], that penalizes abrupt changes in the consecutive differences between coefficients. It is defined for vectors by :

$$\Omega_{S^2} : \boldsymbol{\theta} \in \mathbb{R}^p \mapsto \frac{1}{2}\sum_{i=2}^{p-1}(\theta_{i-1} - 2\theta_i + \theta_{i+1})^2, \tag{3.29}$$

and for matrices, by :

$$\begin{aligned}\Omega_{S^2} : \boldsymbol{A} \in \mathbb{R}^{p,q} \mapsto \quad & \frac{1}{2}\sum_{j=1}^{q}\sum_{i=2}^{p-1}(a_{i-1,j} - 2a_{i,j} + a_{i+1,j})^2 \\ & + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=2}^{q-1}(a_{i,j-1} - 2a_{i,j} + a_{i,j+1})^2 \\ & + \sum_{i=1}^{p-1}\sum_{j=1}^{q-1}(a_{i+1,j+1} - a_{i+1,j} - a_{i,j+1} + a_{i,j})^2. \end{aligned} \tag{3.30}$$

We also include the indicator of the Christmas period in the univariate effects and use the Ridge regularization to penalize the associated coefficient. These univariate features are gathered in Table 3.3 and justified empirically in Section 3.6.4. This setting leads to 366 degrees of freedom for the univariate part in the middle-term model (*i.e.* without the past loads) and 371 in the short-term model.

Name	Category	Symbol	Parametrization
hour of the week	Cyclic	h	168 knots
day of the year		d	128 knots
Christmas period	indicator	1_{xmas}	indicator
timestamp	timestamp	t	linear function
temperatures	acyclic	T^s	17 knots
δ hours delayed temperatures		$\mathsf{T}^s_{-\delta}$	17 knots
last maxima over δ hours		$\bar{\mathsf{T}}^s_{-\delta}$	17 knots
last minima over δ hours		$\underline{\mathsf{T}}^s_{-\delta}$	17 knots
δ hours-delayed load	acyclic	$\ell_{-\delta}$	5 knots

TABLE 3.3: Univariate features for the national forecasts
Set $\mathcal{U}$ of univariate features with the corresponding parametrization for the short-term load forecasting model at the national level.

Bivariate features There are multiple reasons to believe that interactions between the inputs also play a major role for the determination of the load :

- the 2-dimensional conditional expectations observed in Section 2.3.3,
- the good results of (highly non-linear) tree-based models,
- interactions are included in the state-of-the-art GAM models,

- intuitively, a holiday will not have the same effect on working and non-working days of the week, and similarly, the past loads with a fixed delay depend on the day of the week.

While any order interactions can be considered, we did not obtain so far results indicating that explicit interactions between more than two univariate features are useful in a forecasting perspective. In fact interactions between two covariates already introduce an extra layer of complexity. Indeed, as presented in Table 3.4, the number of parameters associated to the unconstrained bivariate features is more than ten times the number of degrees of freedom of the univariate effects. Depending on the quantity of data to fit the model and its regularity which is strongly dependent on the considered level of aggregation, the regularization is essential.

Names	Symbols	Parametrization
cloud covers and day/night	$(\mathsf{c}^s, 1_{\mathsf{sun}})$	3 knots & indicator
week hour and holiday	$(\mathsf{h}, 1_{\mathsf{hld}})$	84 knots & indicator
week hour and day before a holiday	$(\mathsf{h}, 1_{\mathsf{hld}^-})$	84 knots & indicator
week hour and day after a holiday	$(\mathsf{h}, 1_{\mathsf{hld}^+})$	84 knots & indicator
week hour and δ hours-delayed load	$(\mathsf{h}, \ell_{\text{-}\delta})$	84 & 5 knots
week hour and day of the year	(h, d)	168 & 32 knots
temperatures and day of the year	$(\mathsf{T}^s, \mathsf{d})$	5 & 32 knots

TABLE 3.4: Bivariate features for the national forecasts
Set $\mathcal{B}$ of bivariate features with the corresponding parametrization for the short-term load forecasting model.

It is not relevant to impose any of the structures defined in Section 3.3 on interactions between an indicator and another input since the coefficient matrix has only one column or one row. Besides, we have found empirically that the structures on the interaction matrices described in Section 3.3 are not essential for the other inputs so we first consider the unstructured case and discuss this question in Section 3.6.6.

Finally, the interactions including an indicator are regularized with the Ridge penalty $\|\cdot\|_F^2$ of Equation (3.28) while the others are penalized with the smoothing spline regularization Ω_{S^2} defined in Equations (3.29) and (3.30).

Proposed models The version of Equation (3.15) for the middle-term model is :

$$\begin{aligned}
\mathcal{M}_{MT} \quad \sim \quad & \beta_0 + \alpha_1 1_{\mathsf{xmas}} + \gamma_2 \mathsf{t} + f_3(\mathsf{h}) + f_4(\mathsf{d}) \\
& + \sum_{s=1}^{S} \left[f_5^s(\mathsf{T}^s) + \sum_{\delta \in \{24,48\}} \left[f_{6,\text{-}\delta}^s(\mathsf{T}_{\text{-}\delta}^s) + f_{7,\text{-}\delta}^s(\bar{\mathsf{T}}_{\text{-}\delta}^s) + f_{8,\text{-}\delta}^s(\underline{\mathsf{T}}_{\text{-}\delta}^s) \right] \right] \\
& + g_9(\mathsf{h}, \mathsf{d}) + \sum_{s=1}^{S} \left[g_{10}^s(\mathsf{T}^s, \mathsf{d}) + h_{11}^s(\mathsf{c}^s) 1_{\mathsf{sun}} \right] \\
& + h_{12}(\mathsf{h}) 1_{\mathsf{hld}} + h_{13}(\mathsf{h}) 1_{\mathsf{hld}^-} + h_{14}(\mathsf{h}) 1_{\mathsf{hld}^+}. \qquad (3.31)
\end{aligned}$$

The short-term model additionally uses the past loads :

$$\mathcal{M}_{ST} \quad \sim \quad \mathcal{M}_{MT} \quad + \sum_{\delta \in \{24,48\}} \left[f_{15,\text{-}\delta}(\ell_{\text{-}\delta}) + g_{16,\text{-}\delta}(\mathsf{h}, \ell_{\text{-}\delta}) \right]. \tag{3.32}$$

As explained in Section 2.7.1, we are interested in minimizing the NMSE so we define the centered target variable :

$$\mathsf{y} = \frac{\ell}{\bar{\ell}} - 1 \in \mathbb{R}. \tag{3.33}$$

where $\bar{\ell}$ is the empirical expectation of the load ℓ. Thereby, the regularized minimization of the `NMSE` over a training set $[\![1, n]\!]$ is exactly Equation (3.20) where the covariates $(\boldsymbol{x}_i)_{i=1,\dots,n} \in \mathbb{R}^{n,p}$ are obtained by concatenating the defined features.

Finally, when considering the simultaneous forecasting problems of several time series, the corresponding optimization problem is the sum of the individual objectives. It is strictly equivalent with an independent optimization of each problem since there is no coupling in this chapter. Writing this sum allows us to use the general form of Equation (3.25) for the simultaneous optimization of all the models at a given aggregation level. Each target time series has indeed access to its own design matrix since the past loads are not shared and the associated weather stations might be different.

3.5 Experiments with independent models

In this section, we only study the models defined in Section 3.4. We compute in Section 3.5.1 the performances of the middle and short-term forecasts at the different aggregation levels and justify the integration among the inputs of the past loads.

The rest of the section has three goals. First, we illustrate the national model and propose a qualitative analysis to better understand the modeling with univariate and bivariate effects to identify potential difficulties. Secondly, we illustrate the local models to highlight their heterogeneity and the fact that the curves of the estimated local univariate effects are more erratic. Thirdly, we would like to nuance the second point by proving that, in spite of the diversity of the local effects, a similar rough structure is encountered throughout the substations. This last point is particularly relevant to motivate the multi-task setting presented in Chapter 4.

3.5.1 Numerical performances

Collection of training and test sets To assess the performances of the models, we choose the year 2016. Because of the non-stationarity and the obsolescence of data, we have evaluated empirically that, first, the models should be updated every 4 weeks using the new available data, so we consider 4-week-long test sets and secondly, using training sets containing 3 years of data is a reasonable choice. These decisions are discussed in more details in Section 3.7.4.

We measure the numerical performances presented in Section 2.7 with the same frequency as the updates of the models. Therefore, we collect 13 pairs of datasets, similarly to Example 2 of Section 2.7.3. The first pair contains a training set of

exactly 3×52 weeks starting on Friday, January 4$^{\text{th}}$, 2013 and a 4-week-long test set, starting on Friday, January 1$^{\text{st}}$, 2016. The second pair of datasets is obtained by translating the elements of the first pair by 4 weeks forward and so on until the end of the year 2016 is reached. There are consequently 13 pairs of training and test sets.

Relevance of short-term models We compare in Table 3.5 the results of middle and short-term models for the 5 aggregation levels defined in Section 2.5.2. The addition of the past loads among the inputs in the short-term models leads to a substantial improvement at all levels.

	Middle-term	Short-term
National	(0.952, 2.29, 2.89)	(0.983, 1.31, 1.69)
RTE Regions	(0.926, 2.92, 3.68)	(0.959, 2.07, 2.71)
Administrative	(0.925, 3.05, 3.85)	(0.963, 2.07, 2.66)
Districts	(0.910, 3.25, 4.14)	(0.954, 2.27, 2.93)
Substations	(0.842, 7.87, 11.35)	(0.860, 4.64, 7.01)

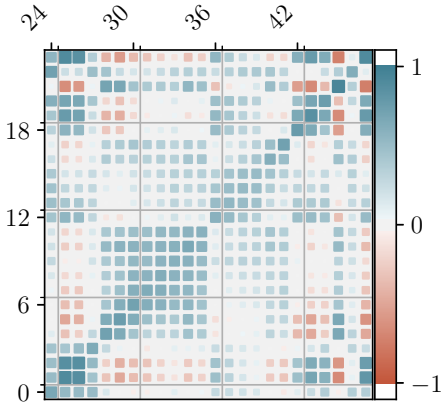
TABLE 3.5: Performances of middle and short-term models ($\texttt{MMr}^2$, MMMAPE, MRMNMSE) as defined in Section 2.7 for the 5 aggregation levels presented in Section 2.5. The short-term models had better results at all aggregation levels on the 13 test sets.

This should not come as a surprise. As explained in Section 3.5 : the past loads may contain complementary information about the current economical activity that is absent from the calendar and meteorological variables. Figure 3.5 shows the correlations between the residuals obtained with a middle-term models and the residuals of the day before. The diagonal from the bottom left corner to the top right corner corresponds to values with a delay of 24 hours. Its non-zero coefficients effectively suggest that short-term models are relevant.

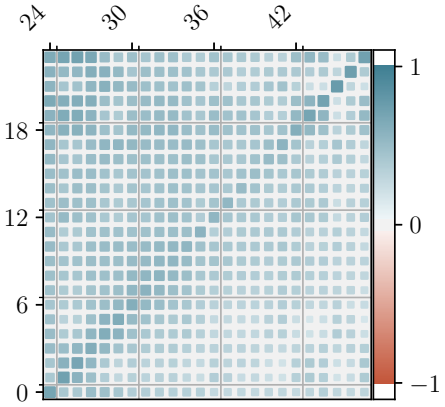
Harder local problems In addition to the expected improvement obtained with the short-term models, there is in Table 3.5 a clear and consistent degradation when the size of the regions decreases for both middle and short-term problems : the smoothing effect due to the aggregation makes the regional and national problems easier than the local problems.

The distribution of the resulting RMNMSE are presented in Figure 3.6. Besides the degradation of the average performance, the disaggregation apparently reveals exotic behaviors that thicken the right queue of the distribution at the level of the substations : complex load series are no longer diluted in the regional sums and represent difficult learning problems.

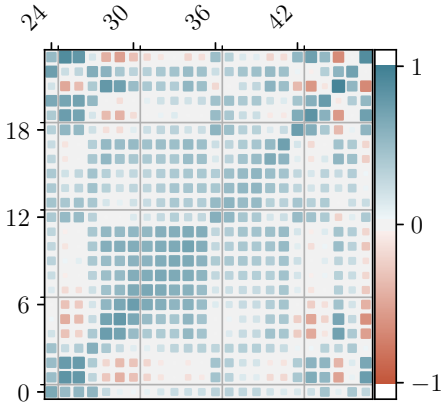
Erratic substations There is also in Table 3.5 a larger difference between middle and short-term models at the substations level. We believe that this is due to a more important impact of the non-stationarity at the level of the substations, that makes the forecasting problems with middle-term models more difficult since they cannot see these variations in the test set while a short-term model may adapt the predictions once the impact of the evolution is caught in the past loads.



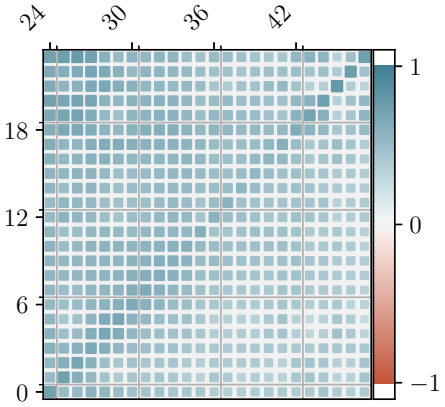
(a) Correlations between the residuals of the national middle-term model on the training set.



(b) Correlations between the residuals of the local middle-term model on the training set.



(c) Correlations between the residuals of the national middle-term model on the test set.



(d) Correlations between the residuals of the local middle-term model on the test set.

FIGURE 3.5: Correlations between consecutive residuals
Correlations between the residuals obtained with middle-term models on day j (hours 0 to 23 on the y-axis) and the residuals on day $j+1$ (hours 24 to 47 on the x-axis), in the training (top) and the test (bottom) sets, at the national (left) and the local levels (right).

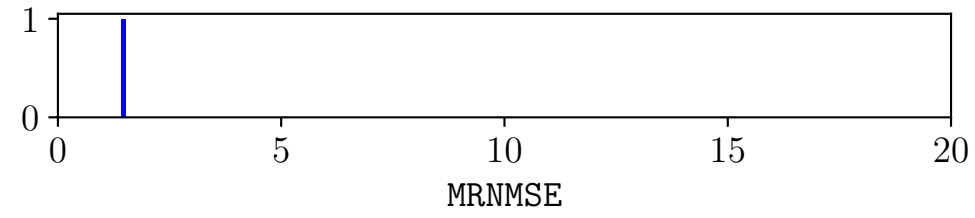
At the substations level, the non-stationarity is more visible for at least three reasons. First, load reports presented in Section B.1 are by nature invisible at the aggregated levels. Secondly, the evolution of the economic and demographic contexts may be relatively faster at this thinner granularity while at the national level, these evolutions are smoothed and slower. Thirdly, the presence of irrelevant values in the substations data may have a dramatic impact on the local predictions while these irrelevant values are smoothed when aggregating the loads, like the variations of the local context.

These problems remain in spite of the data cleansing procedure described in Section B.1. Indeed, we did not make it too restrictive to keep as many substations as

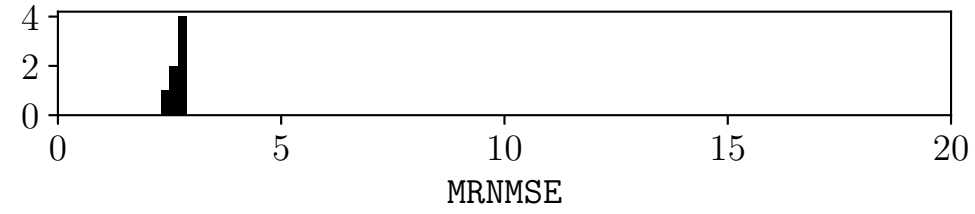
possible, even if some are non-stationary and present jumps related to load reports. Sometimes, it is even difficult to decide visually whether a substation has irrelevant values.

Non-stationary time series are of course difficult to model individually since it can confound the model during the training step or make the numerical performances computed during the test step inaccurate. What's more, including these in a joint modeling for a multi-task learning problem, which is the final goal of this work, might lead to a deterioration of the performances at all substations.

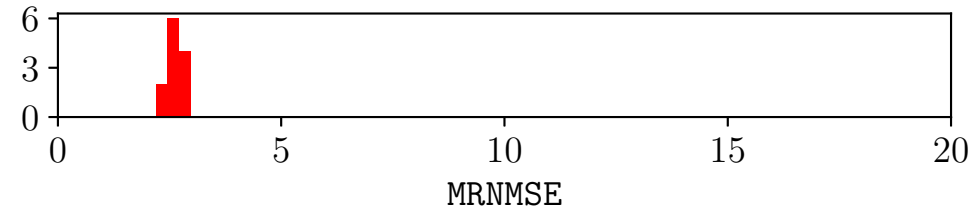
Regarding potential solutions, the evolution of the local context seems unavoidable and requires a more sophisticated modeling. For the load reports and the detection of anomalies, a few methods have been proposed [Jian et al., 2018, and references therein]. Finally, for the irrelevant values, we tried to detect and discard appropriately as much as possible the concerned substations during the cleansing procedure described in Section B.1. However, this problem should be studied further and we clearly believe that identifying these problems before the modeling is a substantial lead for improvement.



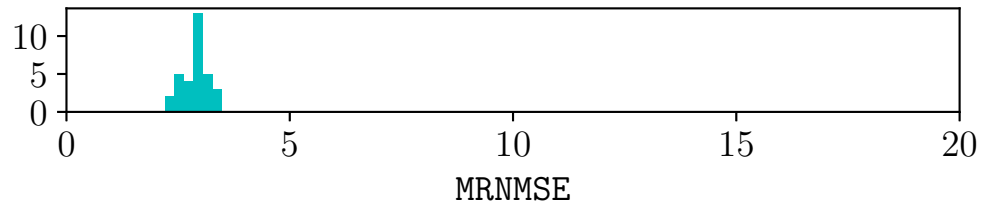
(a) The national load



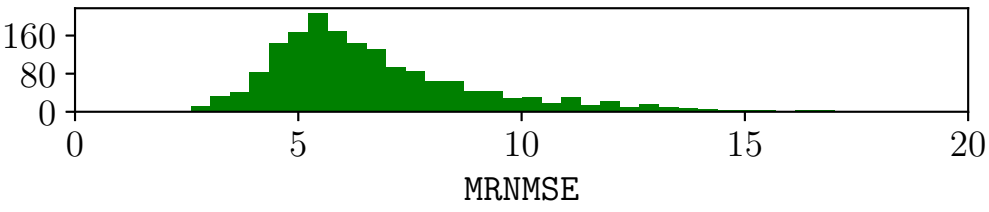
(b) The 7 RTE regions



(c) The 12 metropolitan administrative regions



(d) The 32 districts defined by closest weather stations



(e) The $\mathcal{K}$ substations

FIGURE 3.6: Distribution of the errors
Histograms of the errors $(\texttt{MRNMSE}_k)_{k\in[\![1,K]\!]}$ defined in Equation (2.27), for the 5 aggregation level introduced in Section 2.5.2 by a partition $(Z_k)_{k\in[\![1,K]\!]}$ of the $\mathcal{K}$ substations. Models are repeatedly trained with 3-year-long training sets ending the day before the first observation of the test sets. The RNMSE are computed for each area over 13 batches containing 4 weeks of observations in 2016 and the average is the MRNMSE for each area.

3.5.2 Comparison and validation

We compare the proposed model with the benchmarks at the national and the local levels in 2016 in Table 3.6. The results seem better than the results of the GAM, while the models are quite similar. This might be due to the thinner calibration of the hyperparameters in our model. An interrogation naturally stems from this observation : did the thinner calibration in our models lead to overfitting the test set that is the year 2016 ? To answer this question, we validate our choice of hyperparameters with the year 2017, never seen so far by the models, and compare with the GAM.

	Year 2016	Year 2017
Models of Section 3.4	(0.985, 1.20, 1.59)	(0.951, 1.68, 2.85)
GAM of Section 2.9.4	(0.949, 1.83, 3.16)	(0.931, 1.90, 3.68)
RF of Section 2.9.2	(0.950, 2.12, 2.92)	(0.907, 2.47, 3.92)
XGB of Section 2.9.2	(0.927, 2.81, 3.63)	(0.900, 3.06, 4.21)

(a) National level

	Year 2016	Year 2017
Models of Section 3.4	(0.860, 4.64, 7.01)	(0.849, 4.81, 6.91)
GAM of Section 2.9.4	(0.843, 4.94, 7.47)	(0.831, 5.07, 7.45)
RF of Section 2.9.2	(0.791, 5.66, 8.27)	(0.768, 5.86, 8.44)
XGB of Section 2.9.2	(0.812, 5.60, 7.99)	(0.795, 5.74, 8.07)

(b) Substations level

TABLE 3.6: Validation of the short-term models in 2017
Comparison with the benchmarks on the test year 2016 and the validation year 2017 with ($\texttt{MMr}^2$, $\texttt{MMMAPE}$, $\texttt{MRMNMSE}$) as defined in Section 2.7 at the national and local levels with short-term models. They are learned with the same procedure as for Figure 3.6 but in the right column, the 13 batches of validation are in 2017. The hyperparameters for the Models of Section 3.4 have been selected with the test year 2016 and the year 2017 has never been seen before.

There is indeed a degradation of the $\texttt{MMr}^2$ and the $\texttt{MMMAPE}$ but not always of the $\texttt{MRMNMSE}$ between 2016 and 2017. Plus, our model still outperforms the benchmarks. We cannot say what proportion of this degradation is due to the overfitting and what proportion is due to the non-stationarity.

Remark 5. *We remarked that the optimal choice of hyperparameters evolves and may change between consecutive years. As a consequence, the degradation of the performances between the test year 2016 and the validation year 2017 in Table 3.6 is at least partially due to the fact that the choice of hyperparameters of our models are based on the performances in 2016.*

Put differently, we remarked that exactly the same model can have varying numerical performances over different years with variations up to 20%. This non-stationarity is discussed in more details in Section 3.7.4, although we do not have a definite answer to provide.

3.5.3 Study of the national univariate effects

In order to better understand the model estimated for the **national** load forecasting, we propose in this section to illustrate the univariate effects and the distribution of the residuals. In addition, this allows us to identify potential weaknesses of the chosen modeling. We recall that the national model uses a fictive weather station whose temperature and cloud cover is computed with the weighted mean given in Table F.1.

In the illustrations of this section the model is not updated during 2016 because we fix the training and test datasets to the 3×52 weeks before and the 52 weeks after Friday, January 1$^{\text{st}}$, 2016 so that both are well-balanced and have approximately the same quantity of data for each month, each day of the week and each hour. Thus the distribution of the data in the test set is representative of operational conditions, in terms of possible values of the input variables.

Hour of the week The effect estimated for the hour of the week is presented in Figure 3.7. It is one of the most important effects as its amplitude is the second largest, after the effect of the past loads. The average forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\mathsf{h}]$ are so close on average to the target loads $\mathbb{E}_{\text{train}}[\ell|\mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\ell|\mathsf{h}]$ that the curves are superimposed. Note that the hour of the week is also included in interactions with the day of the year, the past loads and the indicator of holidays.

Every day of the week, the forecasts are less accurate during the day than during the night. Besides, Mondays clearly represent a problem as the residuals are much larger than the other days. We believe that this is because the model includes the effect of the load the two days before and Mondays are the only working days preceded by two non-working days. The inverse situation occurs with transitions from Fridays to Saturdays but it is not as visible in terms of residuals.

Day of the year The effect of the day of the year is presented in Figure 3.8. Although the shape of the learned effect in the national model matches the shape of the load over one year, the amplitude is ten times smaller than the effect of the hour of the week. The residuals are particularly large during the Christmas period and, while they are on average smaller during summer, they are still large during the summer break.

Temperatures The effects estimated for the temperature, the delayed temperature and the extremal values are presented for the national model in Figures 3.9, 3.10 and F.23 - F.27. The norms of the residuals are larger in cold temperatures which typically correspond to larger loads.

Note that the shape of the effect f_5^s in Figure 3.9 learned for the temperature matches the shape of the conditional load but it is not the case of the effects learned for the delayed temperature $f_{6,\text{-}24}^s$ in Figure 3.10 and $f_{6,\text{-}48}^s$ in Figure F.23. However, we should not try and make a causal interpretation of these graphs. Indeed, interpreting the learned effects is not trivial because the inputs are highly correlated.

Also, the amplitudes of the estimated effects related to the temperatures are much larger than the amplitude of the effect of the day of the year presented in Figure 3.8 : according to the learned model, the variations of the amplitude of the

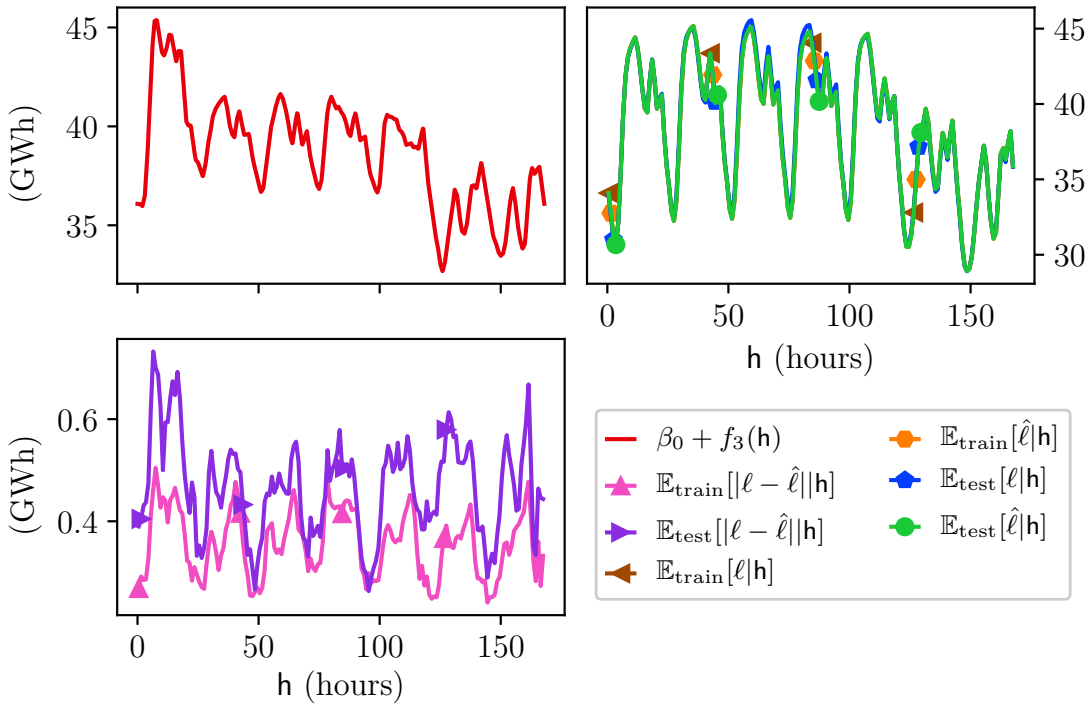


FIGURE 3.7: Estimated effect of the hour of the week
(*top left*) Effect $\beta_0 + f_3(\mathsf{h})$ of the hour of the week estimated in the national short-term model after renormalization.
(*bottom left*) Conditional norms of the residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\mathsf{h}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\mathsf{h}]$.
(*top right*) Target loads $\mathbb{E}_{\text{train}}[\ell|\mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\ell|\mathsf{h}]$ with the forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\mathsf{h}]$.
Note that the residuals are still quite correlated with the hour of the week.

load during the year are much more explained by the changes in the temperatures than by the day of the year.

Past loads The effect $f_{15,\text{-}24}(\ell_{\text{-}24})$ of the 24 hours-delayed loads is presented in Figure 3.11. It seems almost linear, which corroborates the idea that the past load acts like a corrective term, but the slope in Figure 3.11 is slightly smaller than the slope of the empirical loads $\mathbb{E}[\hat{\ell}|\ell_{\text{-}24}]$ that is close to 1. This effect is significant since it has the largest amplitude. The estimated effect $f_{15,\text{-}48}(\ell_{\text{-}48})$ of the 48 hours-delayed loads is presented in Figure F.28.

Indicator of the Christmas period The average values of the national loads, the forecasts and the residual during and outside the Christmas period are given in Table 3.7, for the training and the test sets. Because this indicator is highly correlated with the temperatures and the day of the year, the average values of the different time series are computed only for observations in December and January.

The coefficient α_1 is negative, as expected because the economic activity decreases during this period. However, its amplitude is 5 times smaller than the difference in the training set between the average loads during the Christmas pe-

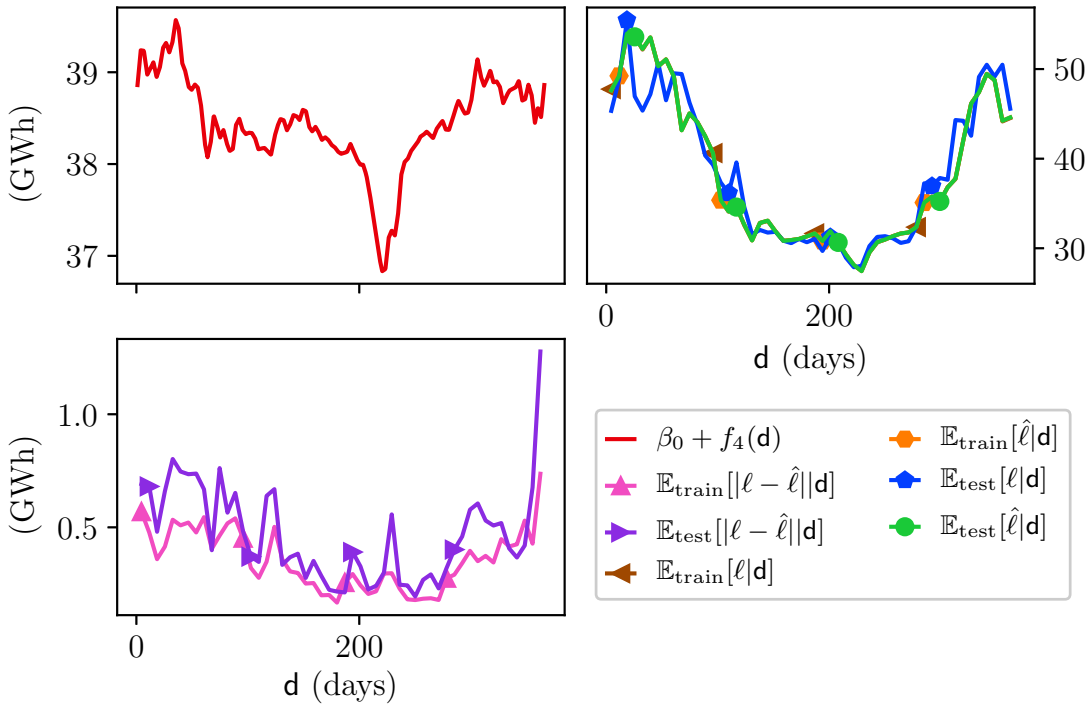


FIGURE 3.8: Estimated effect of the day of the year
(*top left*) Effect $\beta_0 + f_4(\mathsf{d})$ estimated in the national short-term model for the day of the year after renormalization.
(*bottom left*) Smoothed and renormalized graphs of the conditional norms of the residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\mathsf{d}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\mathsf{d}]$.
(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\mathsf{d}]$ and $\mathbb{E}_{\text{test}}[\ell|\mathsf{d}]$ and conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\mathsf{d}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\mathsf{d}]$.
The average loads and the forecasts are very close on the training and the test sets so the curves are superimposed. Note that the important residuals during the Christmas period are not due to boundary effects or to a low density of the data since the day of the year d is uniformly distributed. Also, the learned effect $\beta_0 + f_4(\mathsf{d})$ presents high frequencies. This may be a sign of overfitting, although the hyperparameters have been selected empirically.

riod $\mathbb{E}_{\text{train}}[\ell|1_{\mathsf{xmas}} = 1]$ and the average load $\mathbb{E}_{\text{train}}[\ell|1_{\mathsf{xmas}} = 0, \text{month} \in \{\text{Dec, Jan}\}]$ during the rest of December and January.

Timestamp The last univariate effect in the model depends on the timestamp. In Figure 3.12, we present the smoothed loads, forecasts and residuals over the training and the test sets. Visually, it is not obvious what the linear effect of the timestamp should be. Yet, the coefficient γ_2 in front of the linear timestamp t equals -45 in the national short-term model, which corresponds to a decrease of 45 MWh of the hourly load every year. Still, there was an overestimation of the loads in 2016 since the average residuals are mostly negative. Indeed, if the model can know the test set in advance, that is to say when the model is learned with 4 years of data, the learned coefficient is -49. It remains however a rather small value.

The significativity of these coefficients has not been tested and using the times-

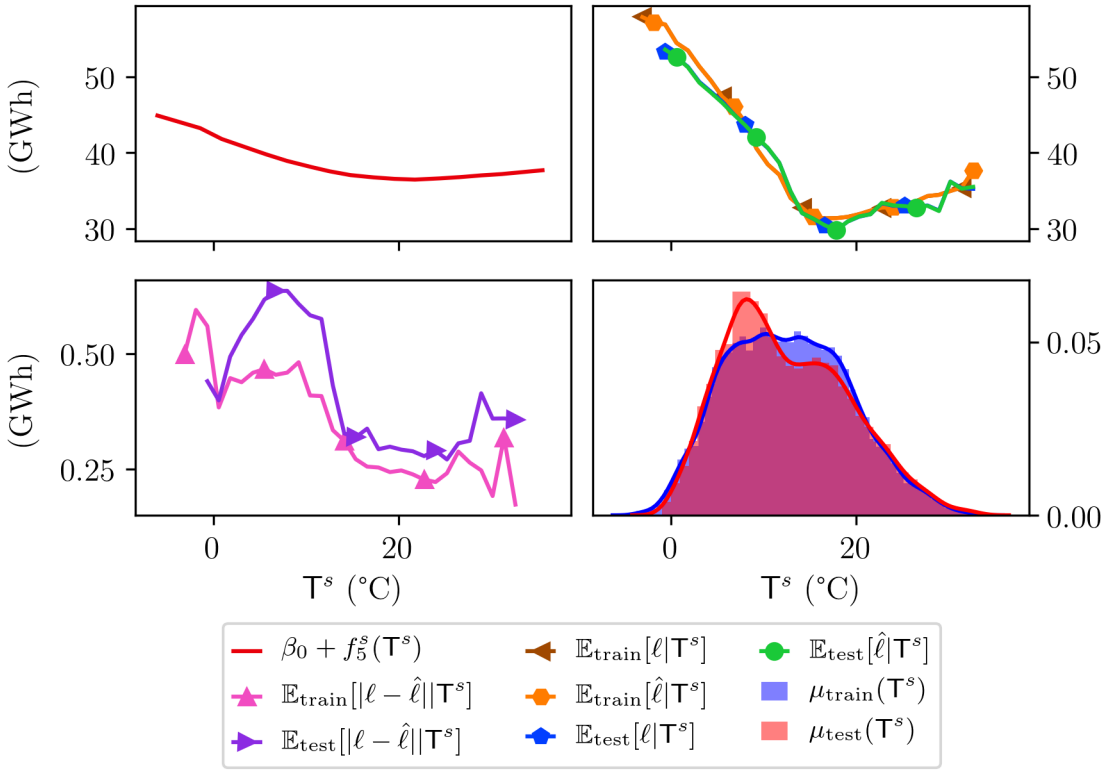


FIGURE 3.9: Estimated effect of the temperature

(*top left*) Effect $\beta_0 + f_5^s(\mathsf{T}^s)$ learned by the national short-term model for the weighted temperature defined in Table F.1 after renormalization.

(*bottom left*) Norm of the conditional residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\mathsf{d}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\mathsf{d}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\mathsf{d}]$ and $\mathbb{E}_{\text{test}}[\ell|\mathsf{d}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\mathsf{d}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\mathsf{d}]$.

(*bottom left*) Density of the data in the training and the test sets.

The illustration of the density of the data shows that the distribution μ_{test} of the temperatures in 2016 was different from the distribution μ_{train} from 2013 to 2015 : this is indeed confirmed by the boxplots in Figure 2.3.

tamp in our experiments only led to a minor improvement. A more refined and efficient treatment of the modeling of the trend over time is proposed by Goude et al. [2013].

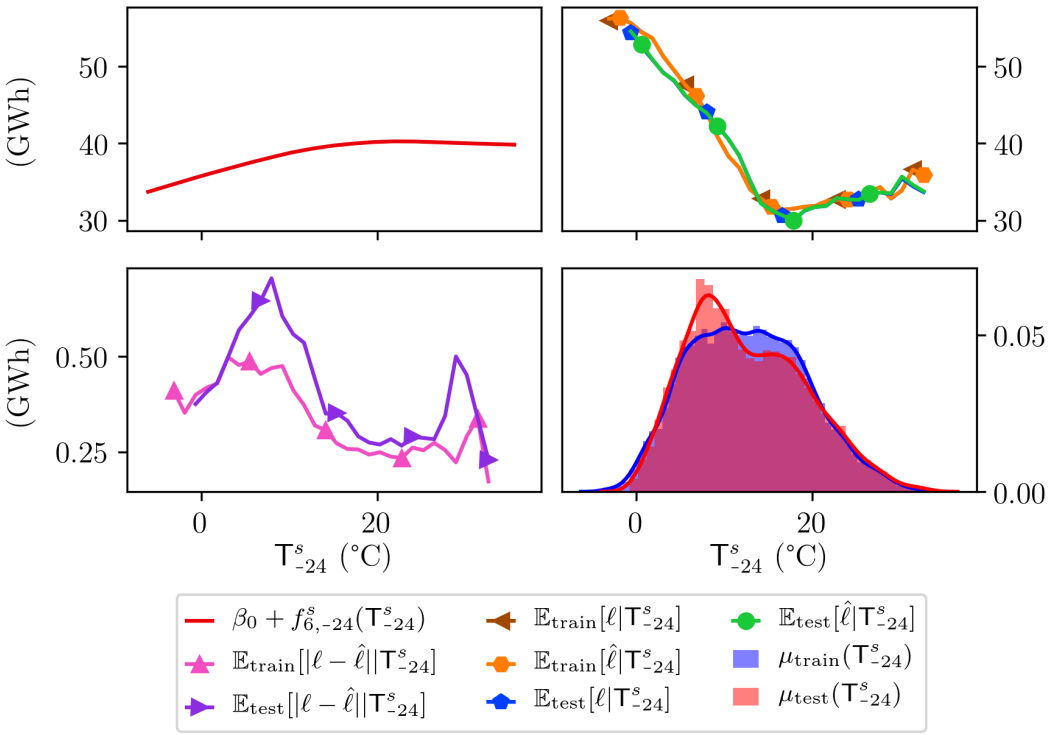


FIGURE 3.10: Estimated effect of the 24 h-delayed temperature

(*top left*) Effect $\beta_0 + f^s_{6,\text{-}24}(T^s_{\text{-}24})$ learned by the national short-term model for the 24 hours-delayed temperature after renormalization.

(*bottom left*) Marginal norm of the residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||T^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||T^s_{\text{-}24}]$.

(*top right*) Target loads $\mathbb{E}_{\text{train}}[\ell|T^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[\ell|T^s_{\text{-}24}]$ with the forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|T^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|T^s_{\text{-}24}]$.

(*bottom right*) Density of the data in the training and the test sets.

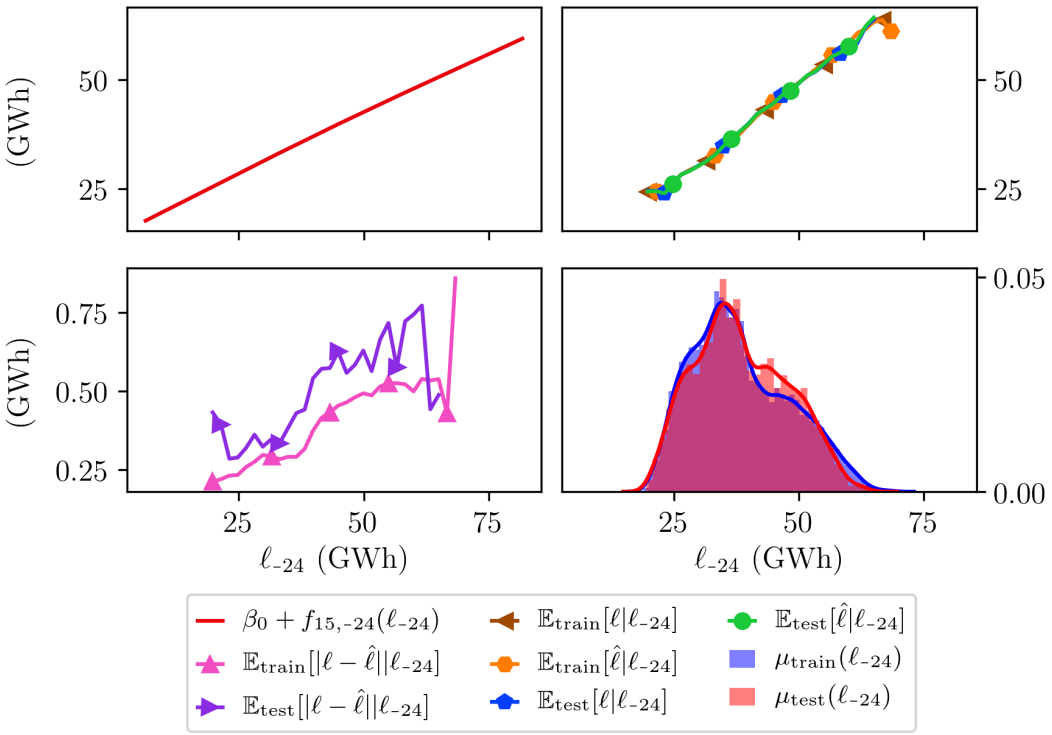


FIGURE 3.11: Estimated effect of 24 h-delayed load

(*top left*) Effect $\beta_0 + f_{15,-24}(\ell_{-24})$ learned by the national short-term model for the 24 hours-delayed loads after renormalization.

(*bottom left*) Conditional norm of the residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\ell_{-24}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\ell_{-24}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\ell_{-24}]$ and $\mathbb{E}_{\text{test}}[\ell|\ell_{-24}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\ell_{-24}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\ell_{-24}]$.

(*bottom right*) Density of the loads in the training and the test sets.

		Christmas period $1_{\text{xmas}} = 1$	Rest of December and January $1_{\text{xmas}} = 0$
Training set	# samples	1152	3240
	Loads (GWh)	44.2	50.7
	Forecasts (GWh)	44.2	50.7
	Residuals (MWh)	-3.54	-25.12
Test set	# samples	384	1056
	Loads (GWh)	46.7	49.8
	Forecasts (GWh)	46.8	49.8
	Residuals (MWh)	-0.087	-0.082
coefficient $\alpha_1 = -0.926$ GWh			

TABLE 3.7: Estimated effect of the Christmas period
Number of samples and average values of the loads, the forecasts and the residuals during the Christmas period ($1_{\text{xmas}} = 1$) and during the rest of December and January ($1_{\text{xmas}} = 0$). The coefficient α_1 is the factor in front of the indicator 1_{xmas} in the national short-term model.

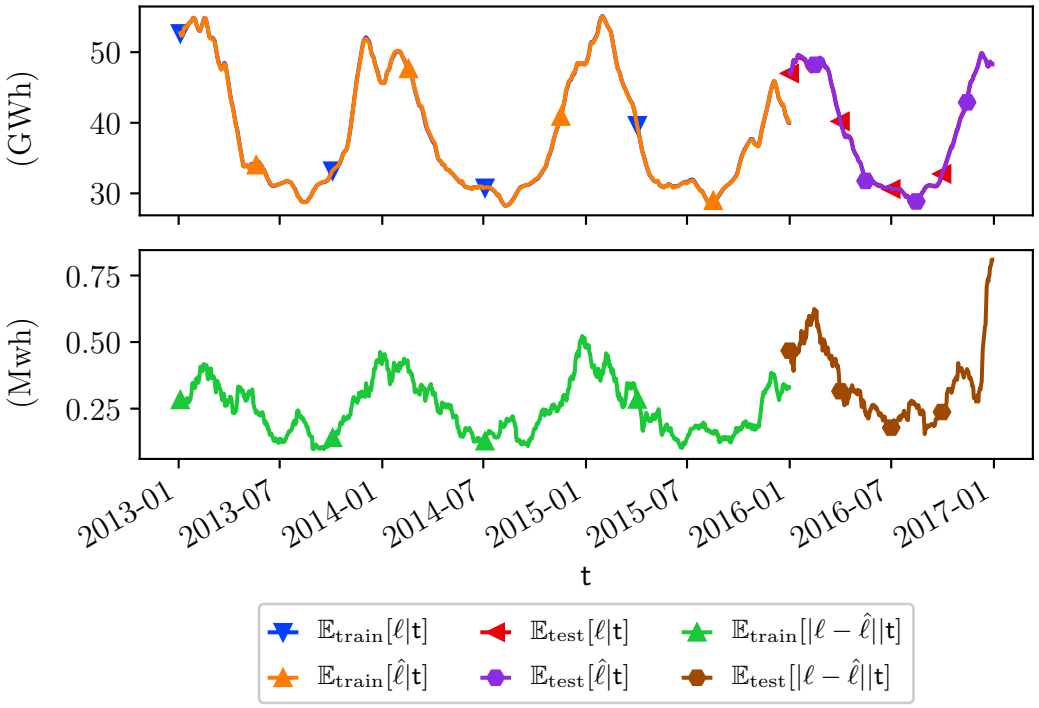


FIGURE 3.12: Forecasts and residuals over the database
(*top*) Loads and forecasts from 2013 to 2015 for the training set and in 2016 for the test set.
(*bottom*) Residuals of the national short-term model over the training and the test sets.

3.5.4 Study of the national bivariate effects

The addition of **interactions** generates more complex models that are more suitable to describe the relationship between the loads and the inputs. In this section, we illustrate the estimation of these relationships in the **national** short-term model, although the interpretations that we can make are less obvious than for the univariate effects.

Past loads and hour of the week It is essential that the effects of the past loads in the model depend on the day of the week since the relationships between the loads of two consecutive days are clearly different if these days are working or non-working days. We consequently included this interaction and it is illustrated for the national model in Figure 3.13. It modifies the effect of the past loads, especially on Mondays ($\mathsf{h} \in [\![0, 23]\!]$), Fridays and Sundays ($\mathsf{h} \in [\![120, 167]\!]$).

Although it is not intuitive why this bivariate effect is small on Sundays and Mondays, what matters to analyze this interaction is the sum $f_3(\mathsf{h}) + f_{15,\text{-}24}(\ell_{\text{-}24}) + g_{16,\text{-}24}(\mathsf{h}, \ell_{\text{-}24})$ which is, on the contrary larger on weekdays than on weekends. This sum is represented in Figure 3.14.

The analogous graphs for the load with a delay of 48 hours are given in Figure F.29 and Figure F.30.

Cloud cover and daylight The estimated effect of the interaction between the cloud cover and the indicator of the daylight is presented in Figure 3.15. Roughly, the more clouds during the day, the larger the electricity demand while this term has no effect during the night ($1_{\mathsf{sun}} = 0$). Note that this effect is relatively small compared with the effects of the temperatures or the hour of the week.

Holidays and hour of the week The impact of the holidays depend obviously on the day of the week. The interactions between the hour of the week and the indicators of holidays are presented in Figures 3.16, F.31 and F.32.

We observe on the learned effect in Figure 3.16 that the electricity demand during working hours from Monday to Friday is reduced on holidays. During the weekend, the estimated effect is almost constant as expected.

In Figure F.31, it appears on the learned effect that the electricity demand is lower on a Monday if it precedes a holiday, which makes sense given that it often leads to a 4-day-long weekend. Whether we see the inverse relation for Fridays in Figure F.32 is debatable. This might be due to the little number of non-working weekdays in the database.

In fact, the effect $1_{\mathsf{hld}^+} h_{14}(\mathsf{h})$ in Figure F.32 leads to larger forecasts on days that follow a holiday. We believe that it is because the potential decrease of the economic activity on days that follow a holiday is already introduced in the model through the effects related to the past loads.

Day of the year with temperature The interaction in the national model between the temperature and the day of the year is presented in Figure F.33. Clearly, the modification induced by the interaction $g^s_{10}(\mathsf{T}^s, \mathsf{d})$ could not be obtained with a sum of univariate functions. However, note that the regions of the input space

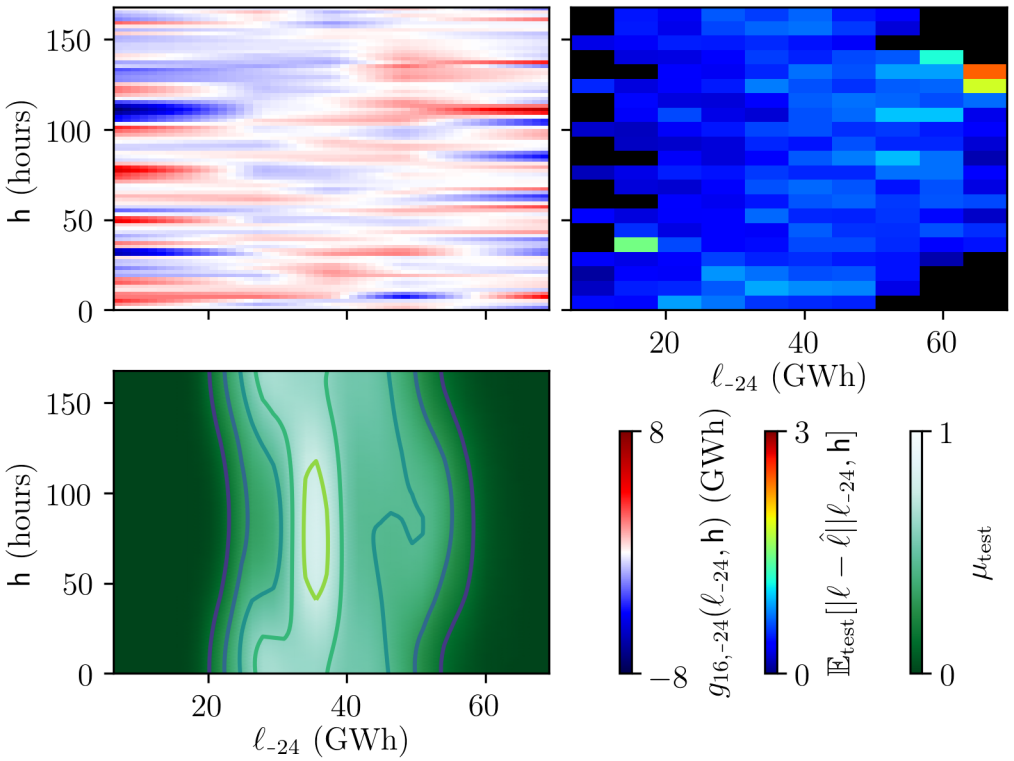


FIGURE 3.13: Interaction between past loads and week hours

(*top left*) Interaction $g_{16,\text{-}24}(\mathsf{h}, \ell_{\text{-}24})$ between the hour of the week and the past load.

(*bottom left*) Estimated marginal density $\mu_{\text{test}}(\mathsf{h}, \ell_{\text{-}24})$ of the inputs in the test set.

(*top right*) Average residuals in the test set $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\mathsf{h}, \ell_{\text{-}24}]$.

where $g^s_{10}(\mathsf{T}^s, \mathsf{d})$ is small correspond to a low-density and their interpretation seems difficult.

Day of the year with hour of the week Finally, the estimated interaction in the national model between the hour of the week and the day of the year is presented in Figure F.34. Although this bivariate effect clearly increases the estimated loads on Mondays during summer and reduces the estimated loads on Saturdays and Sundays during the same period, the coefficients are small compared with other effects.

Also, we can see in Figure F.34 a continuous shift during the year of the peak in the morning from Mondays to Fridays. It occurs around 8 a.m. in summer and one or two hours later in winter. This may be partially due to the Daylight Saving Time (**DST**). However, there is no abrupt transition at the end of March and October, when the shift occurs.

For information, the time in the model and in the plots is always the UTC time and the DST is not explicitly taken into account, any more than with this interaction. Looking closer at electricity load curves near the switch between winter and summer times permits to see a transition but we did not take it into account in the models. We discuss this in more details in Section 3.7.3.

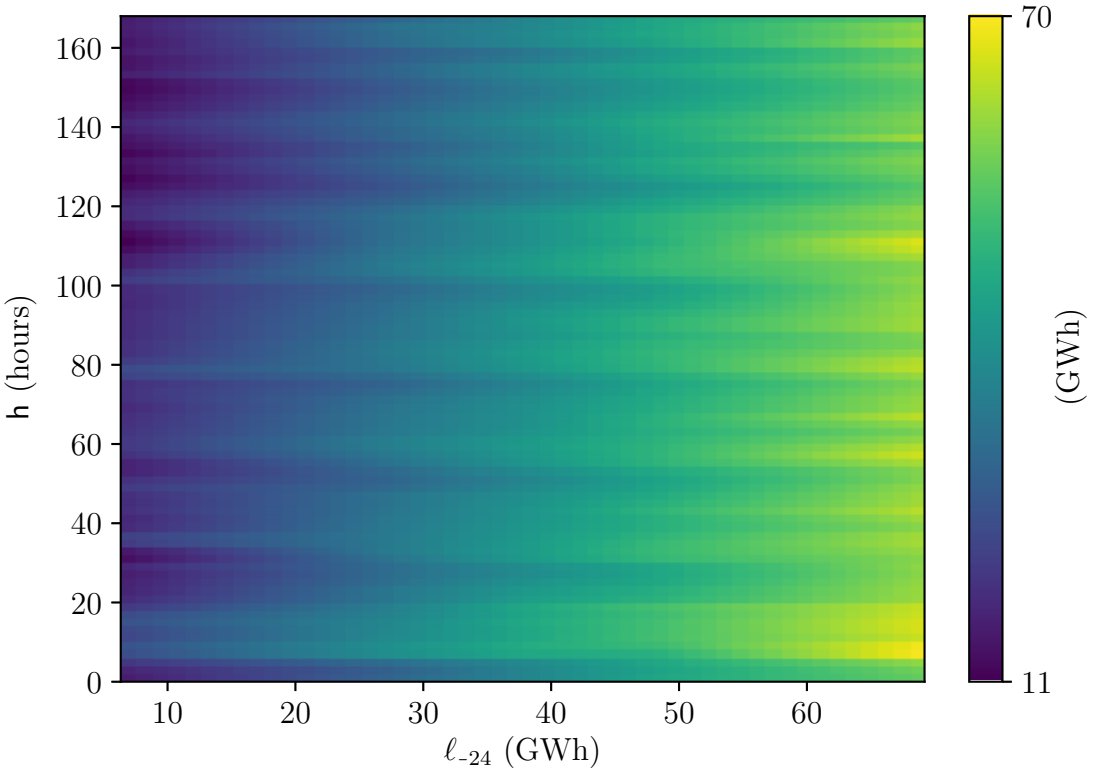


FIGURE 3.14: Total effect of the 24 h-delayed loads
Sum of the univariate and bivariate effects $\beta_0 + f_3(\mathsf{h}) + f_{15,\text{-}24}(\ell_{\text{-}24}) + g_{16,\text{-}24}(\mathsf{h}, \ell_{\text{-}24})$ in the national short-term model.

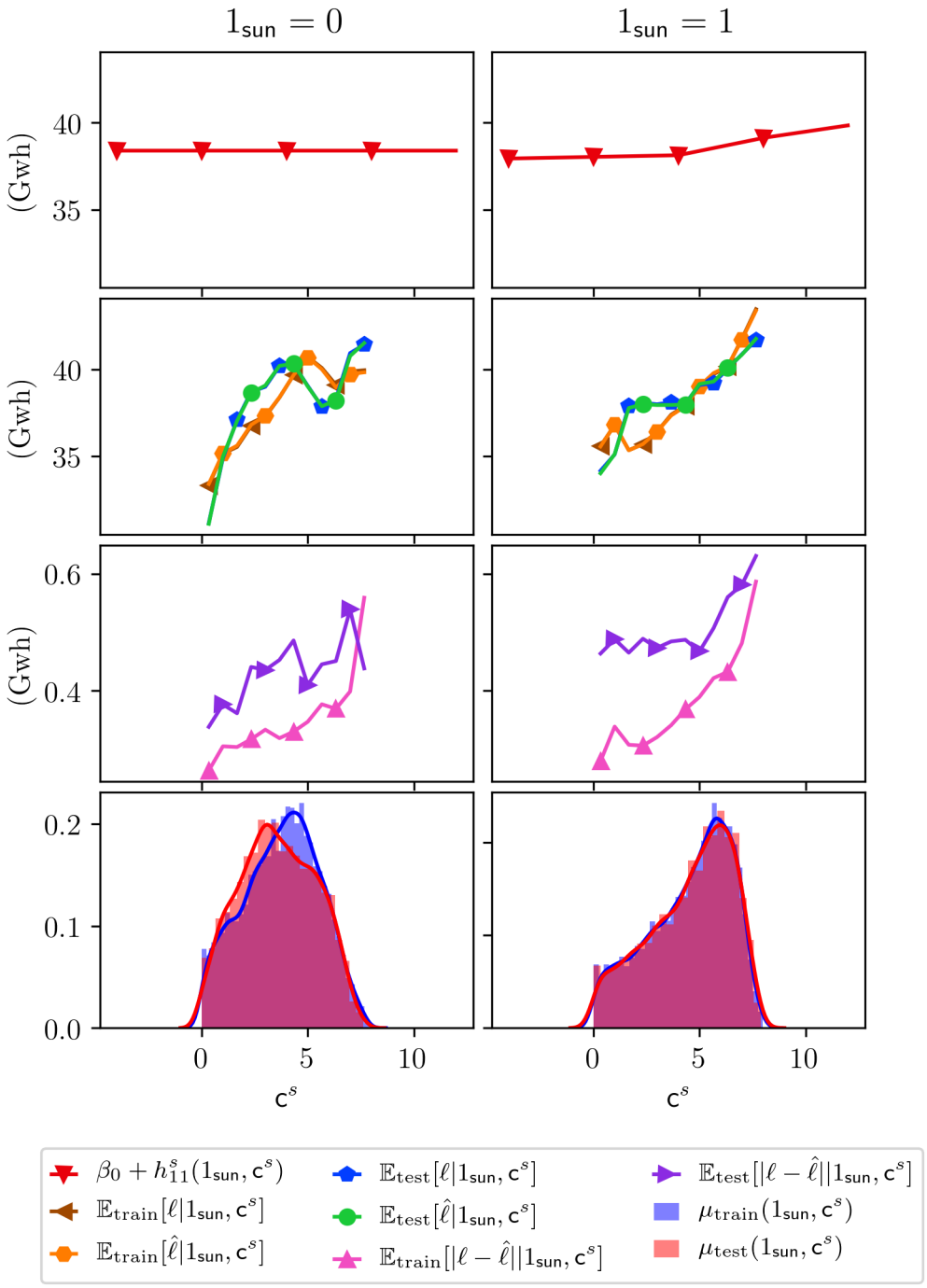


FIGURE 3.15: Estimated effect of the cloud cover during the day

(1^{st} *row*) Effect $\beta_0 + 1_{\text{sun}} h^s_{11}(c^s)$ of the cloud cover, when the sun is down ($1_{\text{sun}} = 0$) and up ($1_{\text{sun}} = 1$).

(2^{nd} *row*) Target loads $\mathbb{E}_{\text{train}}[\ell|1_{\text{sun}}, c^s]$ and $\mathbb{E}_{\text{test}}[\ell|1_{\text{sun}}, c^s]$, with the average forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|1_{\text{sun}}, c^s]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|1_{\text{sun}}, c^s]$.

(3^{rd} *row*) Marginal norm of the residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||1_{\text{sun}}, c^s]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||1_{\text{sun}}, c^s]$.

(4^{th} *row*) Density of the cloud cover in the training and the test sets.

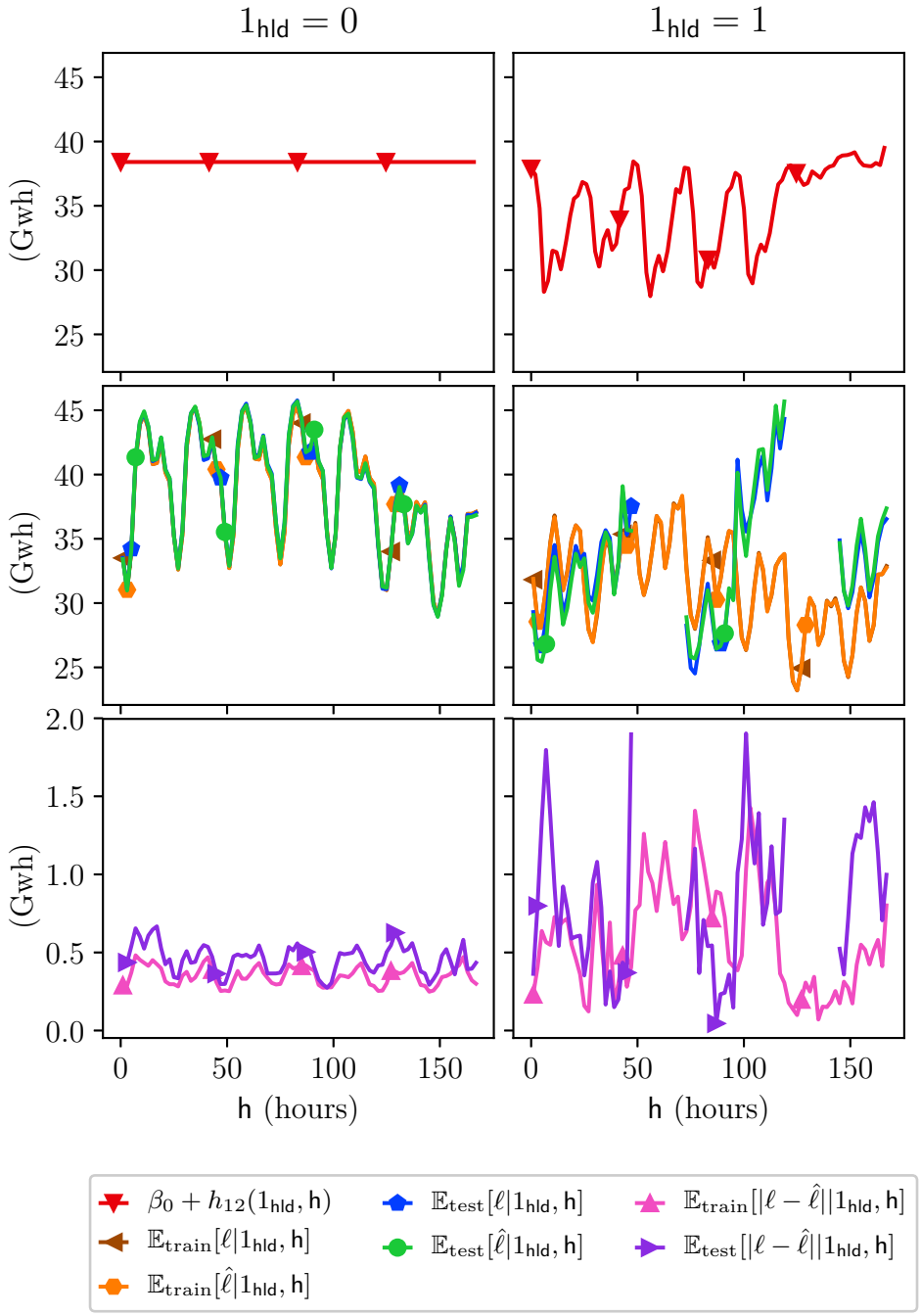


FIGURE 3.16: Interaction of holidays with hours of the week

(1^{st} *row*) Interaction $\beta_0 + 1_{\mathsf{hld}} h_{12}(\mathsf{h})$ between the indicator of holidays and the hour of the week.

(2^{nd} *row*) Target loads $\mathbb{E}_{\text{train}}[\ell|1_{\mathsf{hld}}, \mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\ell|1_{\mathsf{hld}}, \mathsf{h}]$ with the forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|1_{\mathsf{hld}}, \mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|1_{\mathsf{hld}}, \mathsf{h}]$.

(3^{rd} *row*) Marginal norm of the residuals $\mathbb{E}_{\text{train}}[\,|\ell - \hat{\ell}|\,|1_{\mathsf{hld}}, \mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\,|\ell - \hat{\ell}|\,|1_{\mathsf{hld}}, \mathsf{h}]$.

The marginal loads, forecasts and residuals are incomplete in the column $1_{\mathsf{hld}} = 1$ because there was no holiday on Wednesdays and Saturdays in 2016.

3.5.5 Study of the local univariate effects

In this section, we propose a study of the models estimated **at the level of the substations**. The goal is twofold. First a quick look at the distribution of the coefficients for different substations proves that the local loads do not have exactly the same behaviors as the national load and are not perfectly homogeneous. The parameters of the model should consequently be adapted for this level. Secondly, we see that a common structure is nevertheless encountered in these models and it justifies the ambition of models learned within a multi-task framework in Chapter 4.

Since the loads of the substations have different amplitudes, the graphs that we present in this section represent the models estimated to forecast the loads normalized as in Equation (3.33).

Cyclic inputs The remarks about the effects of the hour of the week in the national model are also valid for the substations, namely that the estimated effect of the hour of the week and the residuals are largest on Mondays in Figure 3.17. Marginal quantiles q10 and q90 are relatively close and a similar weekly cycle can be observed for most of the substations.

Similarly, the quantiles of the effect of the day of the year, presented in Figure F.35, show that the quantiles q10 and q90 have a similar shapes, even if this input has a minor effect on the forecasts.

Note that although these illustrations are convenient because they summarize in a single graph the effects estimated for all the substations, they are still limited because they do not guarantee that the effects do not oscillate between quantiles q0 and q100. However, we did observe that for most of the substations, the estimated effects have the same shape as the quantiles and the curves do not oscillate. We do not represent those individual effects to be concise.

Acyclic inputs The quantiles of the estimated effects of the temperature, its delayed and its extremal values over 24 hours are illustrated in Figures 3.18, 3.19, F.36 and F.37. The analogous graphs for a delay of 48 hours are presented in Figures F.38, F.39 and F.40. Again, they exhibit a relationship similar to what is encountered in the national load forecasting model.

Unlike the hour of the week and the day of the year, the temperatures are not cyclic and the fact that quantiles are wide for low and high temperatures is at least partially due to a low density of the data near the boundaries and not only to the fact that load forecasting is more problematic under these weather conditions.

Finally, the quantiles of the effects of the past loads are given in Figure 3.20 with a delay of 24 hours and in Figure F.41 for a delay of 48 hours. Between quantiles q10 and q90, the effect is almost linear in Figure 3.20, like in the national model. Besides, the number of 3 knots has been selected empirically. As explained in Section 3.6.6, the modeling of this input only requires a couple degrees of freedom.

Christmas period In Figure 3.21, we present a histogram of the coefficients learned to model the effect of the Christmas period. They spread around a small negative values. We indeed expect a slight decrease of the loads during this period. However, some substations consistently exhibit an augmentation of the electricity

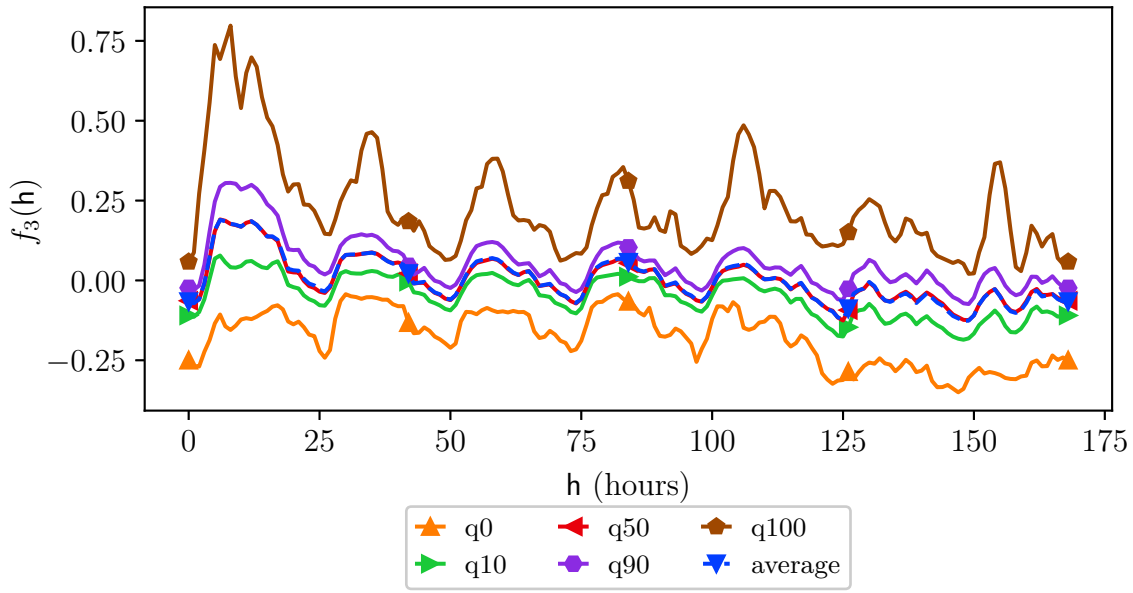


FIGURE 3.17: Local effects of the hour of the week
Quantiles for the estimated univariate effect of the hour of the week at each of the 168 knots. Since the loads at the different substations have different amplitudes the quantiles correspond to the forecasting models for the normalized load, like in Equation (3.33). The mean and the quantile q50 are almost superimposed. Note also that the quantiles q0 and q100 that correspond to the smallest and largest coefficients encountered over the substations are not especially robust.

demand during this period. We believe that this is due to large crowds going to holidays resorts in particular in the Alps, as is illustrated in Figure 3.22. Note however that low temperatures are observed during this period in the Alps and our causal interpretation for the effect of the Christmas period is not certain.

Timestamp A histogram of the coefficients for the timestamp is presented in Figure 3.23. The coefficients spread around zero with both positive and negative values : although the national load seems to be slightly decreasing over time, there might be variations in both directions at a local level due to different local economic and demographic contexts. Note also that the database is quite short to estimate precisely this effect and the significance of the estimated coefficients has not been tested. The apparent decrease could in particular be due to the fact that the electricity demand is smaller during the second half of the year and the database starts in January and ends in December, 5 years later. The introduction of the timestamp in the models is discussed in Section 3.6.5 with Table 3.11.

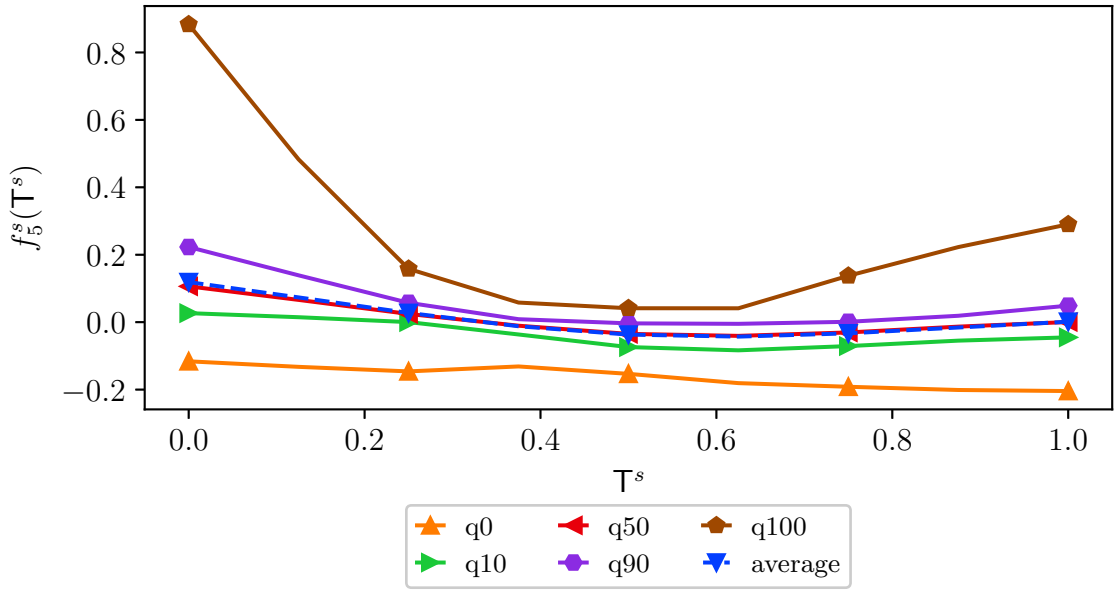


FIGURE 3.18: Local effects of the temperature
Quantiles of the univariate effect associated to the temperature at each of the 9 knots. The mean and the median are almost superimposed. The substations have different amplitudes so the coefficients correspond to the forecasting model of the normalized load. Also, since the substations are associated to different weather stations that have different ranges of possible values, the temperatures that are affinely transformed to lie in $[0, 1]$ have not been transformed back to their original values.

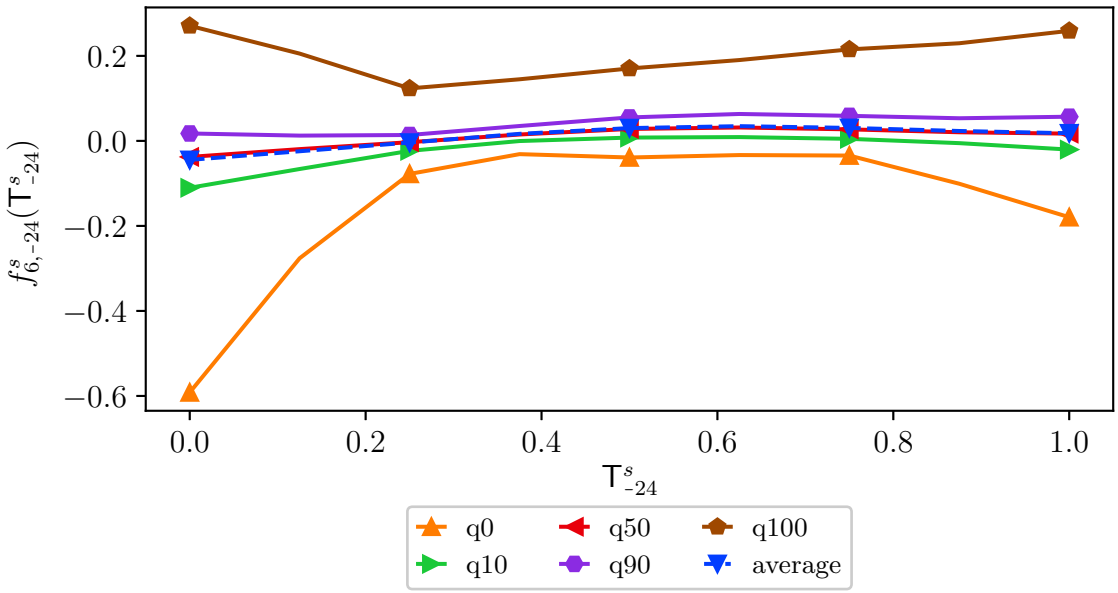


FIGURE 3.19: Local effects of the 24 h-delayed temperatures
Quantiles of the univariate effect associated to the 24 hours-delayed delayed temperatures at each of the 9 knots. The substations have different amplitudes so the coefficients correspond to the forecasting model of the normalized load. Also, since the substations are associated to different weather stations that have different ranges of possible values, the temperatures that are affinely transformed to lie in $[0, 1]$ have not been transformed back to their original values.

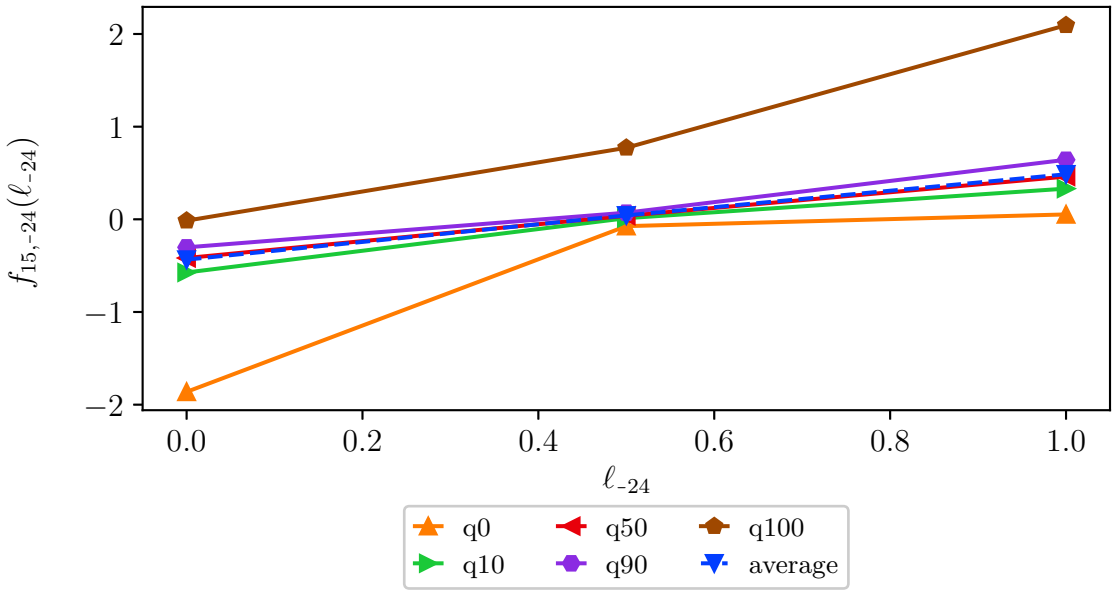


FIGURE 3.20: Local effects of the 24 h-delayed loads
Quantiles of the univariate effects associated to the 24 hours-delayed load at each of the 3 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting models of the normalized loads.

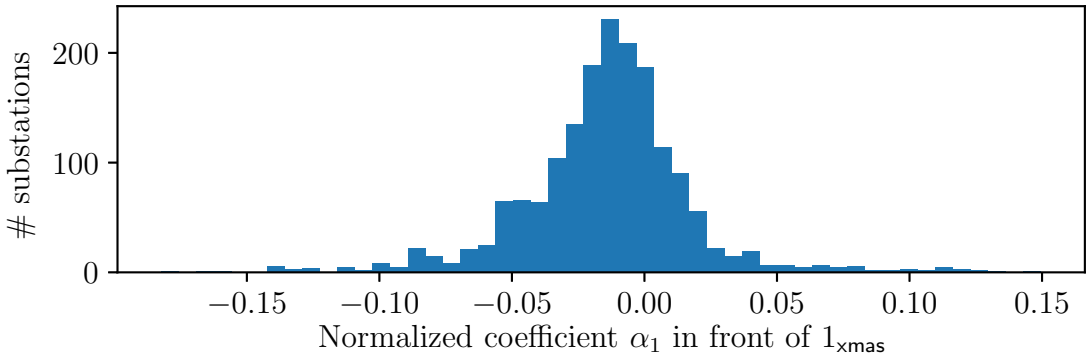


FIGURE 3.21: Local effects of the Christmas period
Distribution of the coefficients corresponding to the Christmas period.

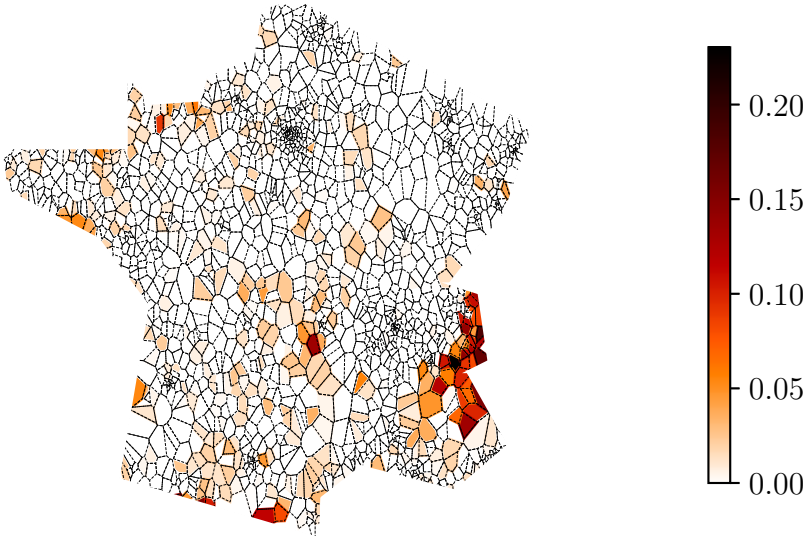


FIGURE 3.22: Positive coefficients for the Christmas period
Positive part $\max(\alpha_1, 0)$ of the coefficients in front of the indicator of the Christmas period for the different substations. Only the positive coefficients were colored. These coefficients are particularly large in the Alps, where many people spend their Christmas holidays.

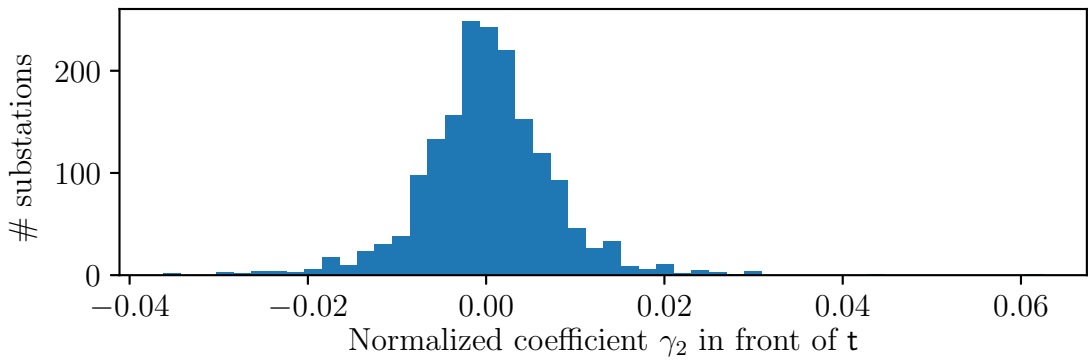


FIGURE 3.23: Local effects of the timestamp
Distribution of the coefficients in the local models corresponding to the timestamp.

3.5.6 Study of the local bivariate effects

In this section, we propose to illustrate the interactions learned by the local models. Given the number of substations and the fact that interactions have large number of degrees of freedom with a lower density of the data to estimate each coefficient than for the univariate effects, we must not undertake a causal interpretation of the graphs and limit the goal of this section to highlight the similarities of the interactions when they are undeniable.

Past loads and hour of the week Since the univariate features of the past loads used for the interaction with the hour of the week are linear functions, these interactions can be represented with quantiles like in Section 3.5.5, in Figure 3.24 for a delay of 24 hours and in Figure F.42 for a delay of 48 hours.

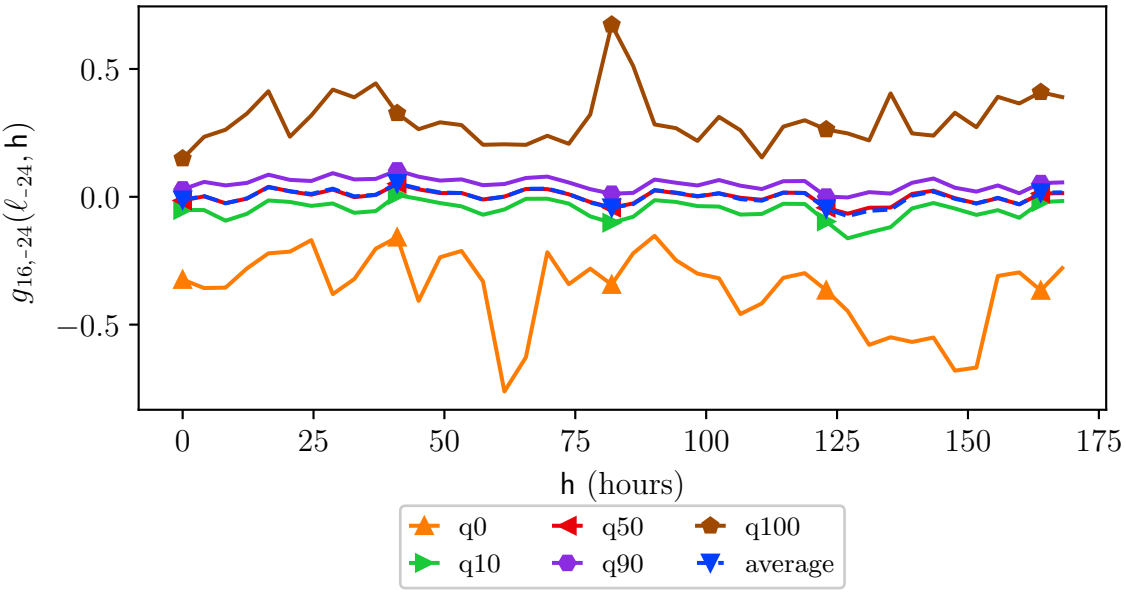


FIGURE 3.24: Local interactions between past loads and hours Given the parametrization of the local models, the features for the past loads used to build the interaction with the hours of the week consist of a single linear function : there exists a function $h_{16,\text{-}24}$ such that $g_{16,\text{-}24}(\ell_{\text{-}24}, \mathsf{h}) = \ell_{\text{-}24} \times h_{16,\text{-}24}(\mathsf{h})$. The quantiles represented on the graph are the quantiles over the substations of the function $h_{16,\text{-}24}(\mathsf{h})$.

Interactions with indicators Similarly, the interactions between the holidays and the hour of the week, as well as the interactions between the cloud covers and the daylight can also be represented with quantiles, because one of the two inputs is an indicator. They are illustrated in Figures F.43 - F.46.

Interactions with the day of the year Finally, to insist on the similarities between the local effects, the standard deviations of the coefficients for the interactions between the day of the year and either the hour of the week or the temperature are given in Figures F.47 and F.48.

3.6 Discussion

From Section 3.1 to Section 3.3, we have defined a standard linear bivariate model. In Section 3.4, we have introduced instances of this generic model specifically designed for day-ahead load forecasting with different variants for the multiple aggregation levels. The resulting models were illustrated in Section 3.5.

The designs of these models are mostly based on the electricity load forecasting literature. However, the local load forecasting problem is relatively new and few papers present a precise setup accompanied by satisfying results at this thinner level. To the best of our knowledge, the best results are obtained with GAM both for the national [Pierrot and Goude, 2011] and the local [Goude et al., 2013] load forecasting problems.

3.6.1 Comments on results

Comparison with benchmarks In short, we believe that the model we introduced gives performances comparable with state-of-the-art models. We implemented the best known benchmarks, with the restrictions mentioned in Section 2.9, and compared the results in Table 3.6.

A significant advantage of the GAM implemented with the MGCV library [Wood and Wood, 2015] is the automatic selection of the hyperparameters : it provides an efficient way of estimating for each substation the best regularization hyperparameters by cross-validation. In order to limit the dimension of the set of hyperparameters at the substations level and to protect these local models against overfitting, we, on the contrary, forced all the substations to use the same hyperparameters and tuned them collectively. We have observed empirically that slightly better performances can be obtained when this constraint is relaxed. However, it makes the search for any further improvement of the model more complex and we choose to maintain this constraint.

Computation Time In terms of computation time, the benchmarks and our model are well under a minute for the prediction of the nationally aggregated load. For the local predictions at the substations level, our implementation, which is based on the quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno (**L-BFGS**) [Liu and Nocedal, 1989; Zhu et al., 1997] implemented in the SciPy library [Jones et al., 2001], leads to a computation time of a couple hours (for one set of hyperparameters). If we only model the univariate effects, this time can be reduced to 15 minutes but the performances are not as good.

Although none of the pioneer papers on local load forecasting presents the speed as a quality of their model, this is a significant gain to test a larger set of configurations. Indeed, Goude et al. [2013] stated that the optimization (with the automatic selection of the regularization hyperparameters) of their local models for 1900 substations takes about about 50 hours. Of course, only the order of magnitude matters since the computers are different.

3.6.2 Main differences with existing models

Spline orders The models that we introduced rely on piecewise linear features while tree-based models rely on piecewise constant functions and GAM are generally piecewise quadratic or cubic. Looking for continuous functions drove us to considering both first and second order piecewise polynomials and we could not conclude empirically that one configuration is better than the other in terms of generalization performances. The sparser structure of the design matrix induced by the first order splines guided our decision since the associated computations are faster.

Hour of the week The treatment of the univariate features in our models is almost identical to the GAM used as a benchmark [Goude et al., 2013; Pierrot and Goude, 2011]. However, we have found empirically that using the hour of the week instead of taking both the hour of the day and the day of the week gives slightly better results, with the Ridge penalty of Equation (3.28) instead of the smoothing spline regularization in Equation (3.29). The way we proceed boils down to considering only the interaction between the hour of the day and the day of the week and never separately. Since it increases the complexity of the model, it can improve the empirical performances only if there are enough regular data, which seems to be the case.

Explicit interactions Because of the results that we obtained empirically, we have included more interactions than the benchmark models. As opposed to the thin plate-splines used by Pierrot and Goude [2011], the interactions defined in Section 3.1.2 have a predetermined number of degrees of freedom. Indeed, each interaction is represented explicitly by a matrix in our model, analyzing them and choosing the appropriate regularization is relatively easy.

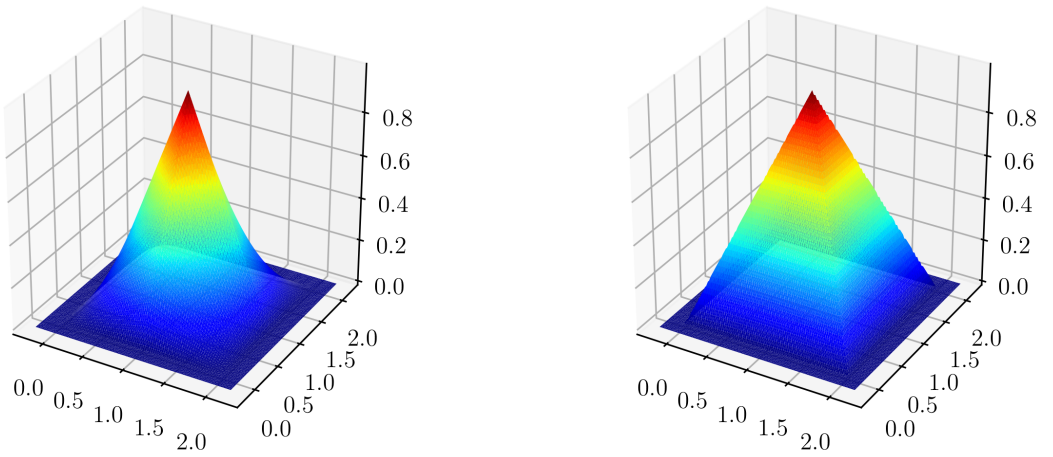
Note that we have introduced the interactions for two inputs $\xi, \zeta \in \mathbb{R}$ as the product $\mathbf{\Phi}(\xi, \zeta) := \boldsymbol{\phi}(\xi)\boldsymbol{\psi}(\zeta)^T \in \mathbb{R}^{p,q}$ in Equation (3.6) of Section 3.1.2. Alternatively, we have considered the interaction $\mathbf{\Phi}'(\xi, \zeta) \in \mathbb{R}^{p,q}$ defined for all $i \in [\![1, p]\!]$, $j \in [\![1, q]\!]$ by :

$$\Phi'_{i,j}(\xi, \zeta) := \min[\phi_i(\xi), \psi_j(\zeta)]. \tag{3.34}$$

Given that the features $\boldsymbol{\phi}(\xi)$ and $\boldsymbol{\psi}(\zeta)$ are piecewise linear, considering $\mathbf{\Phi}'$ leads to piecewise linear interactions, like the univariate features, and would also work for higher order interactions, while the interaction $\mathbf{\Phi}$ leads to second order piecewise functions. Although the empirical results are approximately the same, we believe that these piecewise linear interactions could be considered for future work because they are simpler, easier to represent or manipulate and may lead to more regular functions, as illustrated in Figure 3.25.

3.6.3 Selecting the inputs for the national model

In this section, we justify the choice of the hyperparameters given in Section 3.4 for the **national** model. This setting should not be considered as a definitive optimal parametrization of the models since we observed that over time and depending on the aggregation level, the optimal hyperparameters may vary.



(a) Product interaction of Eq. (3.6) **(b)** Minimum interaction of Eq. (3.34)

FIGURE 3.25: Alternative interactions

Besides, the given choices for the inputs, the interactions, and the granularities, that have discrete values, should be interpreted only as a setting that is close-to-optimal for the database that we have considered. Other works consider different settings and it seems difficult to decide once and for all which choice is the best.

Variable importance In order to justify our choices for the inputs and the associated features in Section 3.4, we propose an ablation study for the national model in Table 3.8. Thereby, we make sure that the introduced covariates are useful for the model, while all the other hyperparameters of the model are kept fixed. In the same table, we also sort the covariates, from the least to the most important for the forecast, in terms of how the performances are deteriorated when it is removed from the model. This should only give an idea of the key inputs for the forecasts but since those are highly correlated, this ranking cannot be considered as the only criterion to consider.

To complete this variable importance analysis, the average norm of the different effects in the predictions at the national level is presented in Table 3.9.

Selection of delayed information The inputs of the model contain information about the loads and the temperatures 24 and 48 hours before the instant to forecast. To choose these delays, we proceeded empirically and present our results for other sets of delays in Table 3.10.

From our point of view, the improvement of the model with past temperatures is related to the thermal masses of the building and possibly to the behavior of people that depends on the recent history. However it is less easy to give a similar interpretation for the past loads.

While the GAM [Goude et al., 2013; Pierrot and Goude, 2011] seem to benefit from having different delays for the different inputs, we use the same 24 and 48 hours delays for all the inputs and did not assess the possibility of having different delays.

	$\texttt{RMNMSE}_{\text{train}}$	$\texttt{RMNMSE}_{\text{test}}$
$\emptyset$	1.29	1.59
Christmas period	1.30	1.60
constant	1.29	1.60
hour of the week	1.30	1.60
day of the year	1.32	1.61
timestamp	1.29	1.61
temperature	1.30	1.62
delayed loads	1.30	1.62
past loads and hour of the week	1.37	1.63
cloud cover and daylight	1.34	1.63
temperatures and day of the year	1.34	1.64
minimum temperatures	1.33	1.64
coming holidays and hour of the week	1.36	1.68
delayed temperatures	1.36	1.69
maximum temperatures	1.43	1.70
past holidays and hour of the week	1.38	1.72
hour of the week and day of the year	1.72	1.84
holidays and hour of the week	2.25	2.77

TABLE 3.8: Ablation study of the national model
Starting from the best national model, we removed the inputs in the left column one by one (with replacement) and evaluated the corresponding performances on the training and the test sets (middle and right column). These inputs have been sorted from the least important to the most important, where importance is defined by the damage the removal of each input does to the `RMNMSE` on the test sets.

	%
$f_{15,\text{-}24}(\ell_{\text{-}24})$	28.90
$f_3(\mathsf{h})$	12.85
$f^s_{7,\text{-}24}(\bar{\mathsf{T}}^s_{\text{-}24})$	8.85
$f^s_5(\mathsf{T}^s)$	8.26
$f^s_{6,\text{-}24}(\mathsf{T}^s_{\text{-}24})$	7.15
$g^s_{10}(\mathsf{T}^s, \mathsf{d})$	5.14
$f_{15,\text{-}48}(\ell_{\text{-}48})$	3.98
$f^s_{7,\text{-}48}(\bar{\mathsf{T}}^s_{\text{-}48})$	3.72
$g_{16,\text{-}24}(\ell_{\text{-}24}, \mathsf{h})$	3.02
$g_{16,\text{-}48}(\ell_{\text{-}48}, \mathsf{h})$	2.92
$g_9(\mathsf{h}, \mathsf{d})$	2.72
$f^s_{6,\text{-}48}(\mathsf{T}^s_{\text{-}48})$	2.43
$f_4(\mathsf{d})$	2.24
β_0	2.03
$f^s_{8,\text{-}48}(\underline{\mathsf{T}}^s_{\text{-}48})$	1.94
$f^s_{8,\text{-}24}(\underline{\mathsf{T}}^s_{\text{-}24})$	1.19
$h^s_{11}(1_{\mathsf{sun}}, \mathsf{c}^s)$	0.84
$h_{12}(1_{\mathsf{hld}}, \mathsf{h})$	0.62
$\gamma_2(\mathsf{t})$	0.46
$h_{14}(1_{\mathsf{hld}^+}, \mathsf{h})$	0.34
$\alpha_1(1_{\mathsf{xmas}})$	0.24
$h_{13}(1_{\mathsf{hld}^-}, \mathsf{h})$	0.15

TABLE 3.9: Average norm of the effects in the national model
Average ratio $100 \times \frac{\mathbb{E}_{\text{train}}[\ |f_d|^2\ ]}{\sum_e \mathbb{E}_{\text{train}}[\ |f_e|^2\ ]}$ for the different inputs over the training set in the national load forecasting model.

Delays (hours)	Performances
(24)	(0.963, 2.0, 2.64)
(24, 48)	(0.965, 1.92, 2.58)
(24, 48, 72)	(0.963, 1.97, 2.62)
(24, 48, 72, 168)	(0.936, 2.48, 3.22)
(24, 48, 168)	(0.927, 2.57, 3.34)
(24, 168)	(0.918, 2.68, 3.47)

TABLE 3.10: Delays for the temperatures and the past loads
Performances in the test year 2016 of the national model ($\texttt{MMr}^2$, MMMAPE, MRMNMSE) as defined in Section 2.7, with different sets of delays.

3.6.4 Tuning the national model

The complexity of the model must be tuned according to the regularity and the quantity of the data in the training set. Once the inputs are fixed, we control it with two sets of hyperparameters : the number of knots and the regularization coefficients. From our experiments, we conclude that the data is more regular at the national level, and we can afford to set relatively large numbers of knots with the appropriate regularizations : the risk of overfitting is low. Therefore, we tend to consider that the most promising leads for improvement at the national level are to consider more inputs, higher order interactions or significantly different models.

As an illustration of the effect of the regularization with the day of the year, and to justify our choice of hyperparameters, we analyze Figure 3.26 and select 128 knots with $\lambda_{\mathtt{d}} = 10^{-3}$, between the small values of $\lambda_{\mathtt{d}}$ where overfitting occurs and the large values where the bias has a negative impact on the RMNMSE. Besides, Figure 3.27 shows the evolution of the effect of the day of the year with 128 knots and penalized by the Ω_{S^2} penalty for different values of the regularization parameters.

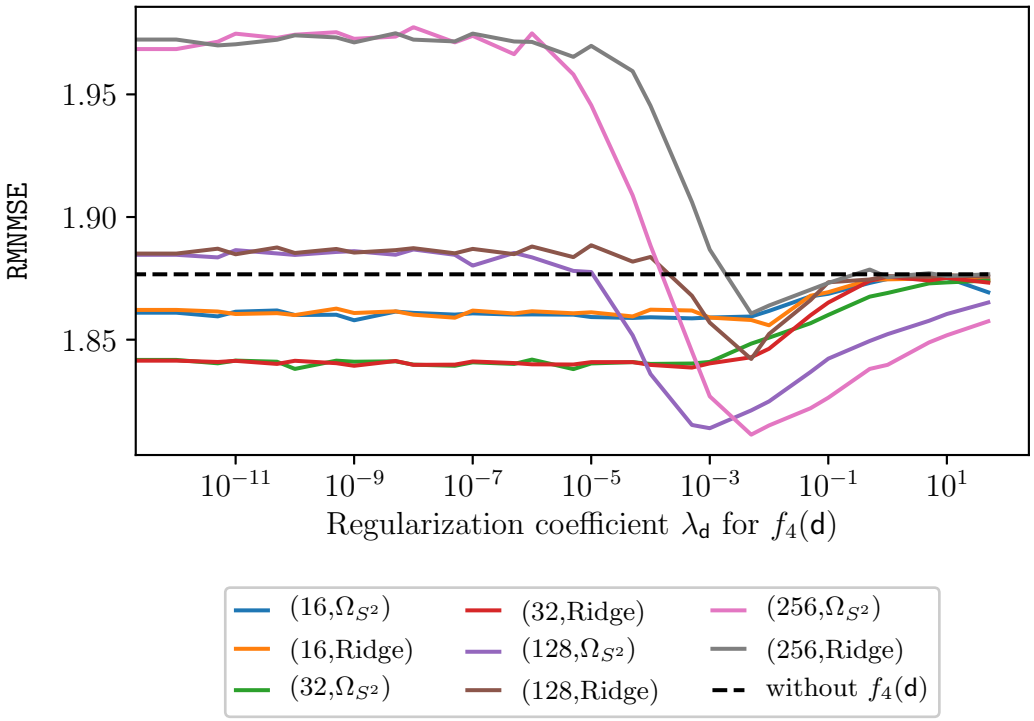


FIGURE 3.26: Regularization of the effect of the day of the year Performances on a test set of different number of knots for the univariate effect of the day of the year. Note the difference obtained with the two possible regularizations : Ridge and the Ω_{S^2} regularization. We selected for the national model 128 knots and the Ω_{S^2} regularization even though the results are slightly better with 256 knots because doubling the number of degrees of freedom is a significant risk of overfitting.

Similarly to Figure 3.27, we present in Figure F.49 the evolution of the univariate effect of the hour of the week for different values of the associated regularization parameter. The performances of the model are not very sensitive to this regularization

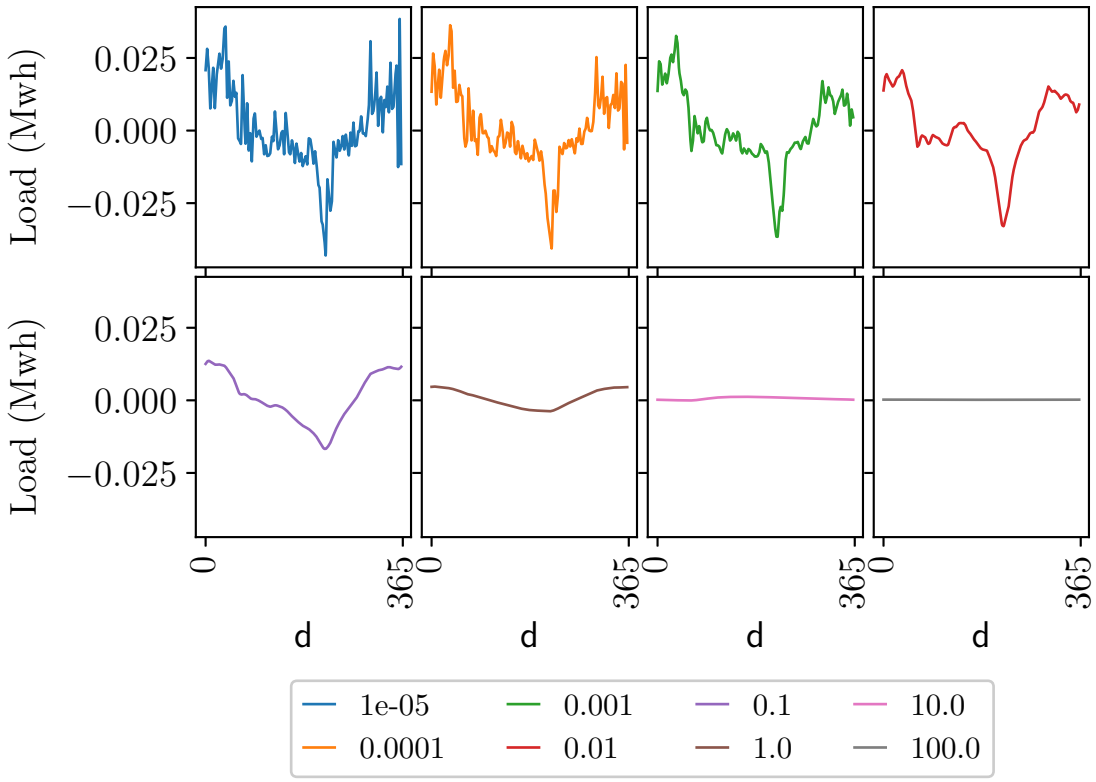


FIGURE 3.27: Evolution of the effect of the day of the year
Evolution of the univariate effect of the day of the year regularized with the Ω_{S^2} penalty for different values of the regularization coefficient. Empirically, we selected 0.001.

parameter. We believe that there are two main reasons. First, the risk of overfitting is low because of the regularity of the data conditioned on the hour of the week. Secondly, and this is probably the most important reason, the hour of the week is also present in interactions and we believe that those can act like substitutes when the regularization coefficient for the univariate features become large. A similar phenomenon is observed with the instantaneous temperatures.

The regularization graphs for the other univariate features are presented in the four Figures F.50 - Figure F.53. Finally, the regularization graphs for the interactions are given in the seven Figures F.54 - Figure F.60.

3.6.5 Selecting the inputs for the local models

For the national short-term forecasting model, it is pretty natural that the only past loads the model can access are past national loads and we did not question this choice. Similarly, we followed the choices made in the literature for the national model about the weather information and used the linear combination of the weather stations given in Table F.1. However, the situation is different for other aggregation levels. We decided that every substation should have access, in addition to the calendar variables, to its own past loads and to the weather information at the 2 closest weather stations. We discuss these decisions in this section.

Choice of the past loads At the substations level, it is legitimate to wonder whether giving each substation access to the past loads of other substations might improve the forecasting performances. The set of possibilities is combinatorial and we had to limit our experiments. We tried to give each substation access to the past national load or to a fixed number of other substations, the 2 geographically closest for example. Empirically, we could not obtain any improvement and concluded that every substation should only be aware of its own past loads.

Number of weather stations In the local load forecasting model of Pierrot and Goude [2011], a meteorologist fixed one weather station for each substation. Since we did not have access to this assignment, we explored two possibilities to try and find an intermediary setting where each substation can access several weather substations but not necessarily all of them, like in the multi-step variable selection procedure proposed by Thouvenot [2015].

First, considering that the most informative weather station for a given substation is not necessarily the closest one, we have tried an automatic selection procedure with a single optimization step : we have given each substation access to all the weather stations but used a group-Lasso penalty so that only a subset is effectively selected. We have also tried a non-convex version of the group-Lasso to reduce the bias [Fan and Li, 2001; Zou, 2006, and references therein].

Secondly, we have given each substation access to a fixed number of the closest weather stations. This provided the best results. According to Figure 3.28, it is best for each substation to have access to the 2 closest weather stations.

Although we conclude here that on average, the local models need access to the 2 closest weather stations to obtain the best performances, it might be because the loads of some substations are in fact more driven by the second closest weather station and not necessarily a combination of the two. It could even be the third closest, but other substations suffer in this case from overfitting. For these reasons, we believe that this question would deserve a longer study.

Ablation study Like for the national model, we present an ablation study of the inputs for the local models in Table 3.11.

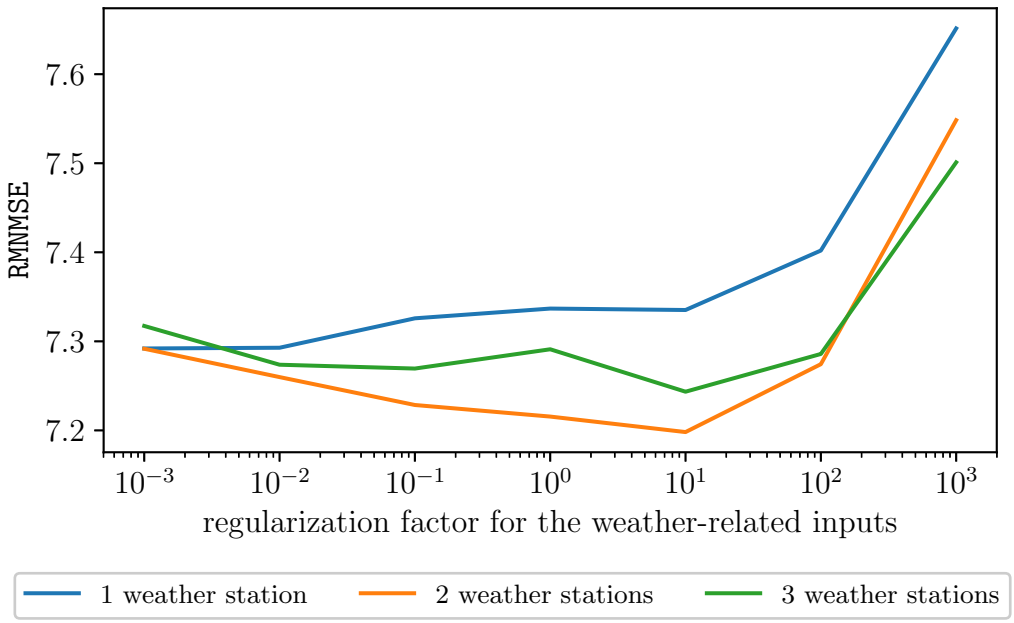


FIGURE 3.28: Selecting the number of weather stations
RMNMSE of the local models for different regularization hyperparameters and numbers of weather stations injected in the inputs. For each number of stations, we compute the value of the RMNMSE on a test set for different values of the weather-related regularization coefficients, where the values are obtained from the best configuration by multiplying them by the same factor. This factor corresponds to the x-axis.
With only one weather station, the models do not overfit the training data but it seems that adding more weather stations makes them more expressive and able to better fit both the training and the test data, with the adequate regularization. The best results are obtained with 2 weather stations.

	$\texttt{RMNMSE}_{\text{train}}$	$\texttt{RMNMSE}_{\text{test}}$
$\emptyset$	5.89	7.22
constant	5.89	7.23
timestamp	5.95	7.25
Christmas period	5.95	7.25
cloud cover and daylight	5.94	7.26
hour of the week and day of the year	6.08	7.28
past loads and hour of the week	5.98	7.29
minimum temperatures	6.04	7.31
day of the year	6.09	7.34
temperatures and day of the year	6.11	7.34
past holidays and hour of the week	6.10	7.35
maximum temperatures	6.16	7.35
coming holidays and hour of the week	6.23	7.43
delayed temperatures	6.13	7.44
temperature	6.16	7.49
hour of the week	6.41	7.52
holidays and hour of the week	6.50	7.62
delayed loads	6.66	8.44

TABLE 3.11: Ablation study of the local models
Presentation of the ablation study with the inputs at the level of the substations. Starting from the best local models, we removed the inputs in the left column one by one (with replacement) and evaluated the corresponding performances on the training and the test sets (middle and right columns). These inputs have been sorted from the least important to the most important, where importance is defined by the damage the removal does to the RMNMSE on the test set. Compared with the ablation study of the national model in Table 3.8, removing the univariate effects of the temperature, the hour of the week or the delayed loads affect much more the local models. On the contrary, removing the bivariate component related to the hour of the week and the day of the year has on average a minor effect on the local models while it is particularly useful for the national model.

3.6.6 Tuning the local models

Contrary to the national aggregation level, overfitting is a more tangible risk at the substations level because the local loads are more erratic. Both the number of knots for each inputs and the regularization hyperparameters must be finely tuned.

Local univariate features We justify, like in Section 3.6.4, our choices of regularization hyperparameters in Figure 3.29 for the hour of the week, in Figure 3.30 for the day of the year and for other inputs in Figures F.61 and F.62.

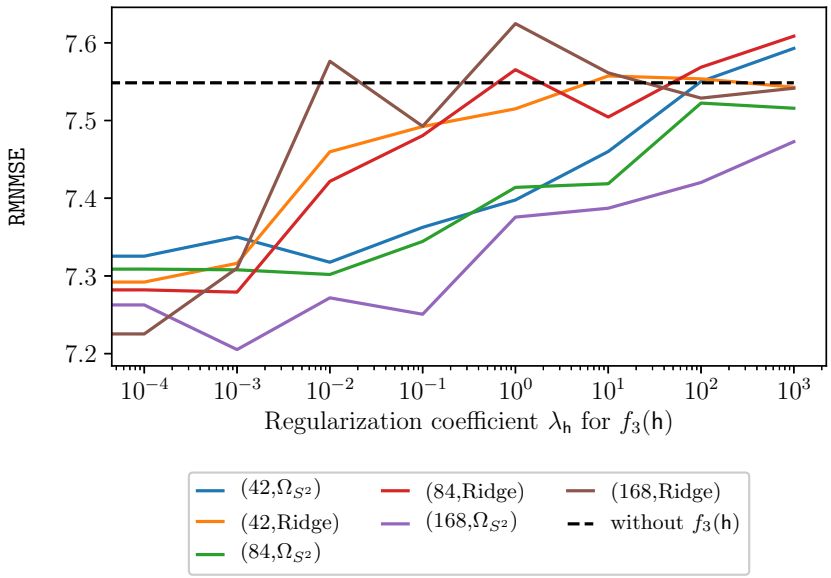


FIGURE 3.29: Regularization for the hour of the week
Performances on a test set with different regularization coefficients and number of knots for the effect of the hour of the week with 2 possible regularizations : Ridge and the Ω_{S^2} regularization.

For the day of the year, we select a smaller number of knots at the level of the substations than at the national level. We believe that this is because we imposed that this number should be the same for all substations and that some substations suffer from overfitting when this number becomes large. Still, we believe that some substations would benefit from a larger number of degrees of freedom and consider that this question would deserve a longer study.

Local bivariate features Like for the univariate features, the complexity of the model induced by the interactions at the level of the substations has to be limited because the risk of overfitting is more tangible, as shown in Figures F.63, F.64 and F.65.

Regularizing the interactions to avoid overfitting has one main drawback : it induces an additional bias since the regularization terms make the objective different from the empirical estimate of the risk. That is why we have introduced in Section 3.3 structural constraints. Indeed, the low-rank constraint of Equation (3.21) and the sesquivariate constraint in Equation (3.23) significantly reduce the number of parameters of the models as well as its complexity.

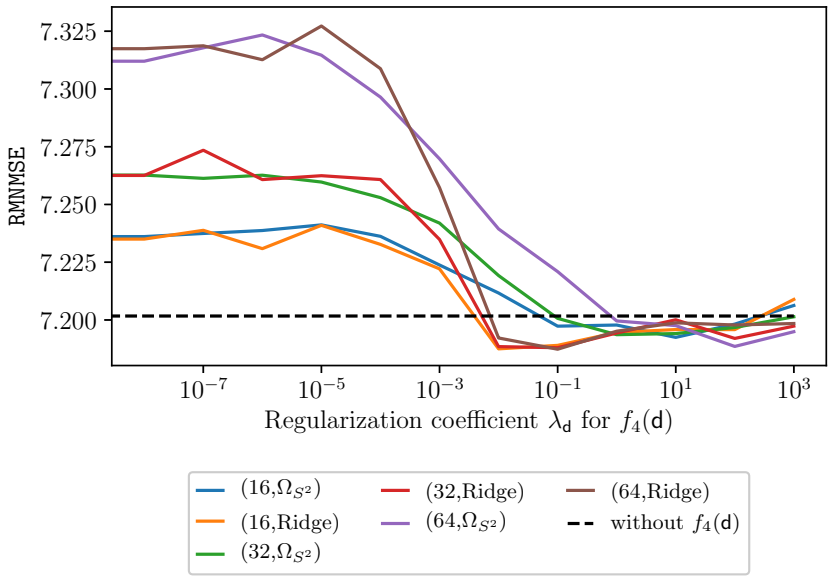


FIGURE 3.30: Regularization of the day of the year
Performances on a test set of different number of knots for effect of the day of the year with 2 possible regularizations : Ridge and the Ω_{S^2} regularization.

Empirically, we have considered using the structural constraint for the interaction between the temperature and the day of the year, the interaction between the day of the year and the hour of the week and the interaction between the past loads and the hour of the week. For the national load forecasting model, these constraints do not improve the generalization performance. At the level of substations, we have found that the low-rank constraint (3.21) does not perform better than the unstructured case while the sesquivariate constraint (3.23) leads to results comparable with the unstructured case.

Thereby, we could not conclude that these structures improve the generalization performance. However, with the sesquivariate constraint that drastically reduces the number of coefficients, we have observed that the convergence is much faster, even though the problem is not necessarily convex. To simplify and keep convex optimization problems, we do not consider these constraints in the rest of this manuscript.

3.7 Pending questions

In this section, we propose to emphasize the problems encountered and the limits of the independent models that we have presented so far.

3.7.1 Important residuals on Mondays

At all levels of aggregation, the short-term models particularly struggle to predict the loads on Mondays, as shown for instance in Figure 3.7. We consider that this is due to the use of the past loads in the model and the fact that Mondays are preceded by non-working days. Although the introduction of the interaction between the past loads and the hour of the week leads to an improvement, it does not solve entirely

the problem. How to deal with sequences of non-working and working days remains unanswered so far, although essential.

3.7.2 Different regularizations for the local models

At disaggregated levels, we chose equal hyperparameters for the different time series to forecast in order to control the dimension of the hyperparameters space during the exploratory part of the work. When trying to fit the best models, this constraint can be relaxed and we have observed empirically that this leads to bettter performances on the test year 2016. However, this relaxation increases the risk of overfitting the training sets and makes more difficult the search for any further improvement of the models. At this point, we do not have a clear answer about the best way to proceed.

Note that by choosing the same hyperparameters for the different models in a same aggregation level, that is to say the same number of knots for the features and the same regularization coefficients, we have somehow already considered the local load forecasting models in a multi-task setting since choosing the best hyperparameters on average let the different models interact together. This is is only a first step towards the more general multi-task settings considered in Chapter 4.

3.7.3 Possibility of additional information

In addition to the delayed temperatures and their extremal values over 24 or 48 hours windows, we have considered different transformations of the weather data. However, we did not obtain any clear improvement by feeding the models with, for instance, univariate features for the cloud cover, exponential smoothing or differences of past temperatures.

To improve the forecasts, we believe that the modeling rather needs other variables. However, in order to simplify the set of inputs and focus on the structure of the models, we have ignored, following expert advice, some extra information that nevertheless are known to marginally impact the demand [RTE, 2014].

Wind and humidity Wind speed and humidity were neglected because they are considered to impact only slightly the electricity demand of end-users. However, the recent evolution of the French electric power system should bring us to question this consideration.

Indeed, the time series that we want to forecast are the demands at the different substations, that correspond roughly to the end-users demand minus the local production, as explained in Section 2.1. Although RTE corrected the load time series in the database to account for the local production of energy that reduces the transit of electricity on the high-voltage network, we know that this procedure might be imperfect (*c.f.* Section 2.1). Therefore, changes of the weather conditions like the wind speed and the intensity of the sunlight might still impact the electricity demand through this mechanism because of the recent development of local renewable energy farms.

Since the sunlight, or more exactly the cloud cover, is part of the inputs of our model, we can hope that the local solar production is automatically taken into

account. However, wind speed is not part of the data that we use. Put differently, wind speed may not have a considerable impact on the demand of end-users but probably has one on the load at the substations level. We believe that this question requires further investigation.

Substations for national forecasting For the national load forecasting problem, the historical model of EDF has access to a fixed weighted mean of the conditions at 32 weather stations [Pierrot and Goude, 2011], as explained in Section 2.5.2. This choice of this meteorological information was not thoroughly questioned.

In order to better understand what this weighted mean is, we illustrate in Figure 3.31 its coordinates in the 2-dimensional space spanned by the 2 first principal components of the matrix whose columns are the temperatures at the different weather stations, after its rows have been centered.

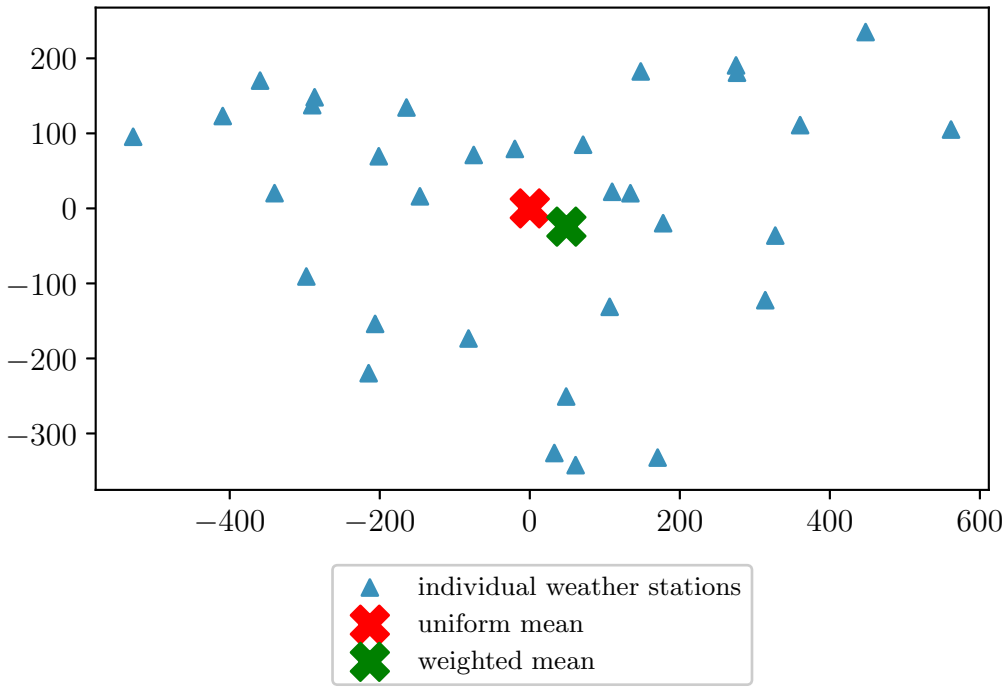


FIGURE 3.31: Projection of stations on 2 principal components Projection on the 2 principal components of the centered temperature matrix, of the time series at different weather stations. The national uniform mean and the weighted mean (*c.f.* Table F.1) are also plotted with crosses.

Besides, we assess in Figures F.66 and F.67 the possibility of summarizing the information at the 32 weather stations in a lower dimension vector with the best rank-r approximation of the temperature matrix.

In the national model, we could include the first principal components instead of the weighted average to enrich the information about the weather. Alternatively, we tried to find automatically a linear combination during the optimization of the model, with and without variable selection penalties. However, none of our limited experiments let us conclude that another combination of the weather stations could

lead to better numerical performances at the national level. Therefore, we have kept the weighted average of the 32 weather stations for the national model.

Weather forecasts for local models At the local level, we feed the models with the weather at the 2 closest weather stations. Even though this number was selected empirically, other settings should be considered.

In particular, Météo-France provides richer forecasts, at approximatively 4000 points of a grid covering the French territory. Using forecasts with this thinner granularity can potentially improve the quality of the load forecasts, as long as the relevant information is selected for each local model.

Special tariffs In France, some of EDF customers benefit from a reduced price on electricity most days of the year in exchange for a high tariff on some (cold) days, where the electricity demand is particularly high and expensive for the producers. These are the so-called *special-tariff* contracts [RTE, 2014]. The high-price days are announced the day before and logically impact the demand [Thouvenot, 2015, Figure 2.3]. These days are made public on the Eco2Mix website [RTE, 2019a] but we have ignored those to simplify the modeling.

At least, these days can be discarded when evaluating the numerical performances of the model, as done by Pierrot and Goude [2011]. We have tried this and since the relative order of the models remained the same, all the results in this manuscript include those days. Other demand-response mechanisms, specific to the electricity markets [Wikipedia, 2019], are not public but potentially impact significantly the demand too.

Holidays In addition to the features related to the day of the year, we have introduced in the load forecasting model an indicator of the Christmas period to take into account the notable decrease of the economic activity during this period. We have considered using a similar indicator for the summer break because of the specificity of this period, from mid-July to the end of August but could not conclude that it helped the models.

Given the improvement with the introduction of the Christmas period indicator at the level of substations, we believe that taking into account other vacation periods (winter, Easter, autumn) can potentially also improve the forecasts. However, these vacations are not nationally synchronized and their introduction into the model requires a more refined treatment.

Ignored Daylight Saving Time Every collected data was measured and injected in the models hourly with UTC time. Thus, shift between Standard Time (**ST**) and Daylight Saving Time (**DST**) are ignored during the measuring process. They are also ignored in the modeling so far.

While it is difficult to measure how this change of regime impacts the forecasts when it is not taken into account, we noticed the shift of the peaks in the morning and in the evening in Figures 2.14 and 2.9. Besides, we illustrate the national load demand near the shifts between ST and DST in Figure 3.32 and while we notice differences before and after the shift, it is not clear how this issue should be dealt with.

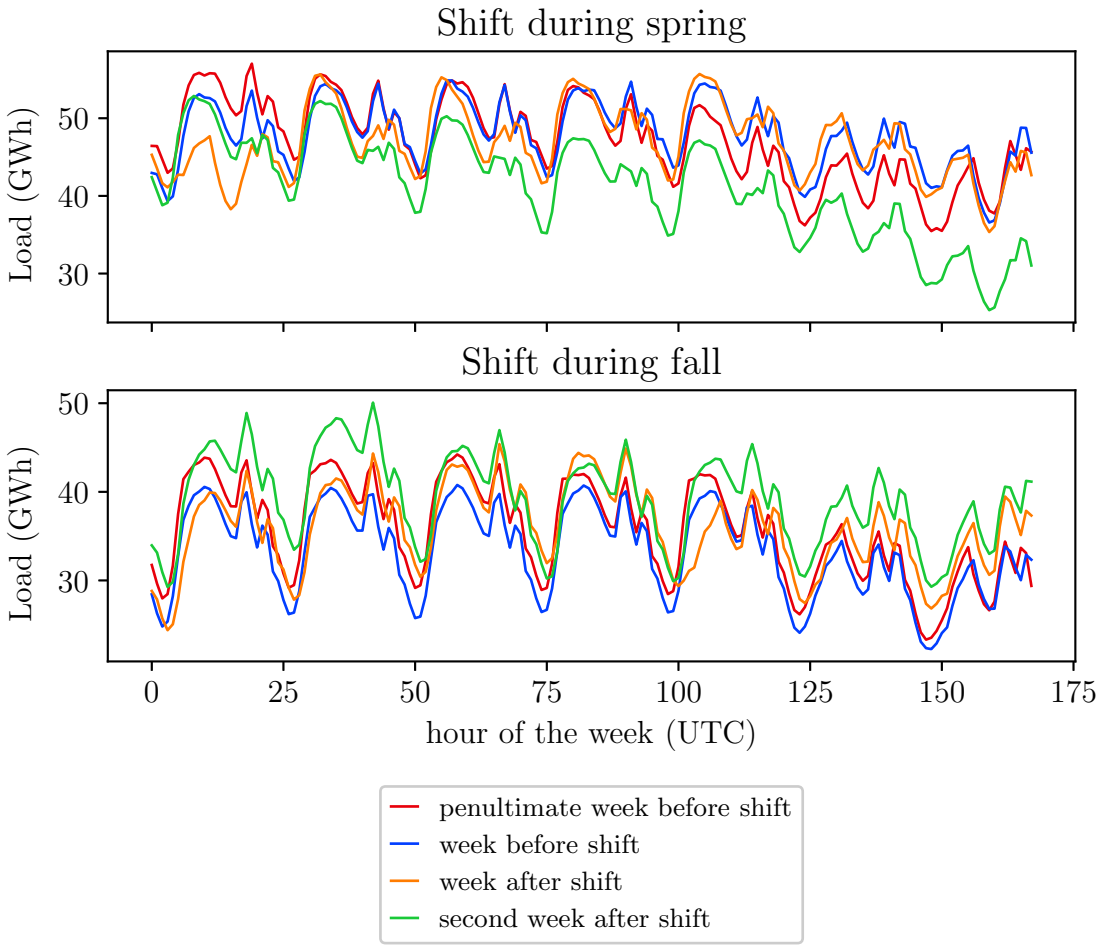


FIGURE 3.32: Effect of Daylight Saving Time
National loads measured with the UTC time during the weeks in 2013 near the shifts between Standard Time (**ST**) and Daylight Saving Time (**DST**) in spring and fall. The shift occurs on Sunday morning (hours 144 to 168) during the *week before shift* (blue). While the peaks during spring seem to occur earlier on the yellow and the green curves (after the shift *i.e.* DST) and earlier on the red and blue curves during fall (before the shift *i.e.* DST), there is not a clear translation of 1 hour of the curves. Besides, the weather conditions are not constant during these months, and the time of the sunrise and the sunset evolve fast. A precise study would certainly require to measure the loads with a smaller time step (in minutes instead of hours).

Higher order interactions Finally, we limited our study to interactions between pairs of inputs (considering that the hour of the week is a single input). However, the effect of holidays, that depends on the day of the week, as shown in Figure 3.16, probably also depends on the day of the year and this is not possible in our models.

We could consider higher order interactions with three inputs. The number of triplets is combinatorial but they may increase the expressiveness of the model and potentially helps fitting the data better. Note that they are possibly taken into account by the highly non-linear tree-based models described in Section 2.9.2.

3.7.4 Dataset shift

The electricity demand is a non-stationary time series. As a consequence, it is not true that, the bigger the training set, the better the estimated model. In this section, we identify some of the problems induced by the non-stationarity and present potential remedies.

Dataset shift In order to describe the problems related to non-stationarity, consider the abstract problem of forecasting at different times $t \in \mathbb{R}$ a target variable $\mathsf{y}_t \in \mathbb{R}$ with a vector of covariates x_t and an estimator $\hat{\mathsf{y}}_t = f_\theta(\mathsf{x}_t)$. Let ℓ denote an arbitrary loss function, $\mathbb{P}_{\text{train}}$ and $\mathbb{P}_{\text{test}}$ denote the joint distribution of $(\mathsf{x}_t, \mathsf{y}_t)$ respectively in the training set and in the test set. For simplicity, we assume that these distributions are constant in each set.

The final objective is the minimization of $\mathbb{E}_{\mathbb{P}_{\text{test}}}[\ell(\mathsf{y}_t, f_\theta(\mathsf{x}_t))]$ and we estimate a forecasting model by solving the empirical version of the following minimization problem (regularization aside) :

$$\min_{\theta} \quad \mathbb{E}_{\mathbb{P}_{\text{train}}}[\ell(\mathsf{y}_t, f_\theta(\mathsf{x}_t))]. \tag{3.35}$$

For this procedure to succeed, $\mathbb{P}_{\text{train}}$ and $\mathbb{P}_{\text{test}}$ must be as close as possible. **Dataset shift** is precisely the situation where $\mathbb{P}_{\text{train}}$ and $\mathbb{P}_{\text{test}}$ are different. It designates the variation over time of the joint distribution that can be written with Bayes' theorem :

$$\mathbb{P}_t(\mathsf{x}_t, \mathsf{y}_t) = \mathbb{P}_t(\mathsf{x}_t)\mathbb{P}_t(\mathsf{y}_t|\mathsf{x}_t). \tag{3.36}$$

In forecasting problems, we distinguish two subcases based on Equation (3.36) :

- **Covariate shift** [Sugiyama and Kawanabe, 2012] is the evolution over time of the distribution $\mathbb{P}_t(\mathsf{x}_t)$,
- **Concept drift** [Gama et al., 2014] is the evolution over time of $\mathbb{P}_t(\mathsf{y}_t|\mathsf{x}_t)$.

Alternatively, these subcases can be linked with Equation (3.35) via the law of total expectation :

$$\mathbb{E}_t\left[\ell(\mathsf{y}_t, f_\theta(\mathsf{x}_t))\right] = \mathbb{E}_{\mathsf{x}_t}\left[\mathbb{E}_{\mathsf{y}_t|\mathsf{x}_t}\left[\ell(\mathsf{y}_t, f_\theta(\mathsf{x}_t))|\mathsf{x}_t\right]\right]. \tag{3.37}$$

Both of these problems are essential for load forecasting since they have an impact on the optimal length of the training sets and on the frequency of the model updates, in particular at disaggregated levels where the non-stationarity is more visible.

Covariate shift In operational conditions, the problem we are interested in only consists in forecasting the loads for the next day, and ideally the forecasting model is estimated the day before. This is a reasonable assumption since the model that we study are estimated in a couple hours.

To learn the model, the training set should represent as well as possible the day to forecast. That is in particular why we have considered well-balanced training and test sets in terms of hours of the weeks, since they both contained whole weeks. However, the problem is more intricate for other inputs. Although, this

problem concerns multiple inputs, we focus in this section on the temperatures, as the interpretation appears simpler.

The training set should contain data samples with temperatures similar to the day to forecast. Roughly put, to predict a day in January, we should mainly consider in the training set data measured in January, possibly in different years. Immediately, this leads us to consider a discontinuous training set but this approach has two noticeable drawbacks. First, parametrizing the form of the history increases the number of hyperparameters. Secondly, this may lead to a smaller training set while, originally, we would like as much data as possible. Besides, we could not significantly improve the results with such a parametrization of the training history.

Alternatively, importance sampling is a promising lead. It consists in weighting the data in the training set in order to mimic the distribution of the data in the test set. Let $\hat{\rho}(\mathsf{x}_t)$ denote an estimator of the ratio of the densities of $\mathbb{P}_{\text{test}}(\mathsf{x}_t)$ and $\mathbb{P}_{\text{train}}(\mathsf{x}_t)$. A solution to the loss minimization over the test set can be estimated by solving the empirical version of :

$$\min_{\theta} \quad \mathbb{E}_{\mathbb{P}_{\text{train}}}[\hat{\rho}(\mathsf{x}_t)\ell(\mathsf{y}_t, f_\theta(\mathsf{x}_t))]. \tag{3.38}$$

Although the density estimation problem is difficult in high dimensions, a sequence of work is precisely dedicated to the estimation of ratio of densities [Sugiyama and Kawanabe, 2012; Sugiyama et al., 2012].

While the differences of the marginal densities of the inputs can easily be illustrated, their impact on the numerical performances of the forecasting models cannot be measured. Unfortunately, we could not significantly improve the results with such weighting scheme but we consider that this problem demands a longer effort.

Concept drift The operational problem is inductive : given a training set, we must extrapolate the available information, Past and Present, to forecast the Future. Since concept drift induces the obsolescence of data, old observations only carry little information about the current relationships between the loads and the covariates.

At the national level, it seems that concept drift is sufficiently weak so that using the longest history (possible within the 5 years dataset) leads to the best results but at disaggregated levels, our conclusions are different. Therefore, we cannot consider that the more data we have to train the models, the better. In order to forecast non-stationary time series, we have considered 2 patches.

First, the length of the training set should depend on the considered aggregation level. Even between areas within the same aggregation level, empirical results indicate that the optimal length of the training set can vary. It can even vary over time. We present in Figure 3.33 the numerical performances of the short-term local models for different history lengths and for different frequencies of the model updates in Figure 3.34.

Secondly, the addition of the timestamp among the inputs may help modeling the concept drift and empirically, it leads to better results as presented in Tables 3.8 and 3.11 where we see that the results are not as good when the timestamp is removed from the inputs. A more efficient treatment is the *detrending procedure* presented by Goude et al. [2013, Section II-C] that significantly improves their results. Note that more exactly, their procedure accounts for the evolution over time of $\mathbb{P}_t(\mathsf{y}_t)$.

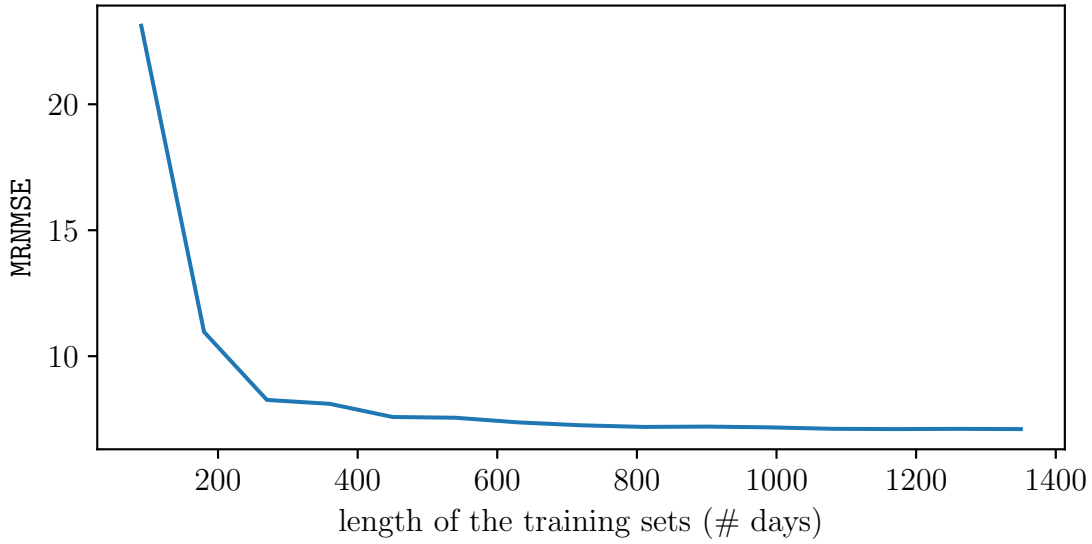


FIGURE 3.33: Results for different lengths of the training set
The local models are tested with the year 2017 for different lengths of the training set. In the rest of the manuscript, we use 3-year-long training sets.

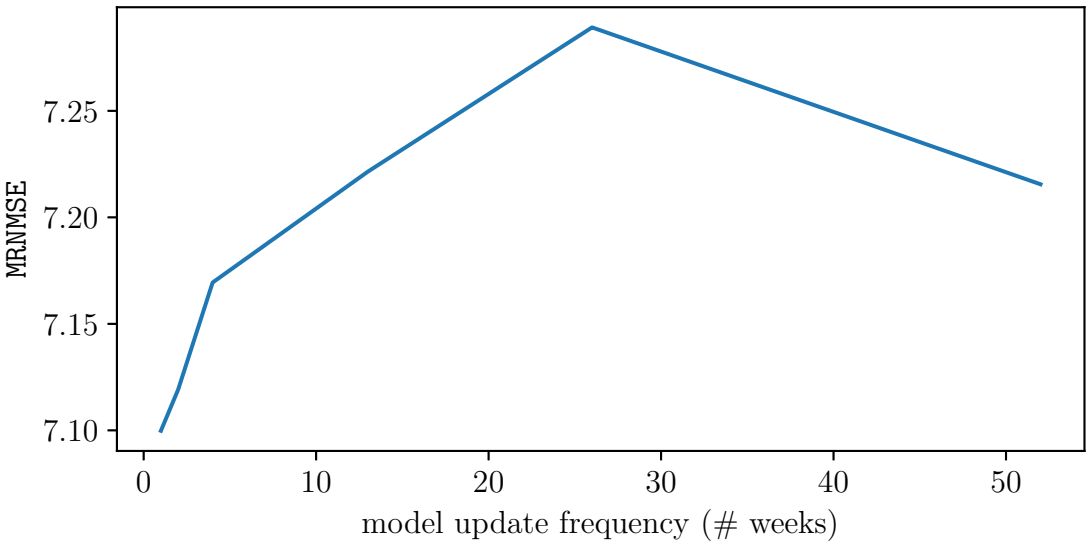


FIGURE 3.34: Model updates frequency
MRNMSE for the local models with different frequencies to update the model in 2016. The frequency is the same for all the substations and in the rest of this manuscript, we update the model every 2 weeks.

Difficult marriage of covariate shift with concept drift. Given a day to forecast, we would like on the one hand to have data as recent as possible because of data obsolescence and on the other hand, we would also like to have data that corresponds to similar conditions *i.e.* where all the inputs lie in a part of the input space similar to the day to forecast. These two requirements are difficult to fulfill jointly and at this point we consider that for the load forecasting problems :

1. Selecting empirically the most appropriate length of the training set like in Figure 3.33 is the first priority.
2. Integrating the detrending procedure of Goude et al. [2013] represents a low-cost patch to concept drift in terms of implementation, and should be evaluated.
3. Studying further the possibility of improving the model with importance sampling may lead to potentially significant improvements.

3.7.5 Choice of the numerical criteria

Finally, we have introduced in Section 2.7.1 different possible numerical criteria to optimize and assess the load forecasting models. The criteria to optimize the models and the criteria to assess them should be the same but it is not clear to domain-experts which one it should be. In fact, there is in the load forecasting literature a contradiction as many models are optimized using squared errors while they are commonly compared with the `MAPE`.

In our work, we have mainly focused on the `RMNMSE` for both optimization and comparison of models, justifying in Section 2.7.1 its potential interest for the TSO. Nevertheless, we believe that the numerical measure could be different if we take into account for the local forecasts all the operational constraints and costs induced by the errors for the TSO.

3.8 Conclusion of Chapter 3

In the beginning of this chapter, we have introduced a standard bivariate linear model for load forecasting. We have detailed the parametrization at the national and the local levels, particularly emphasizing the risk of overfitting and the importance of regularizing appropriately the different univariate and bivariate effects.

The analysis of the local predictions let us identify the difficulties of local load forecasting resulting, among others, from the non-stationarity of the electricity demand. At all aggregation levels, we have identified Mondays as an important problem, due to their occurence after non-working days. While we admit that generic Machine Learning models may not be the best tool to forecast the load during isolated events like holidays, solar eclipses and World Cup Finals, we should be able to deal with Mondays, the summer break and the Christmas period. In particular, the way the past loads are introduced in the models is relatively basic and a more sophisticated approach might be relevant.

At the level of substations, the more erratic behavior of the electricity demand makes the forecasting problems more difficult. The aberrant values in the database are clearly an important obstacle but even with a better procedure to clean the database, there would remain jumps in the time series due to load reports. The ability to detect these seems to be a key step before proposing more accurate load forecasting model.

Finally, we have tried to emphasize the similarities between the local models but how to measure this similarity properly and detect outliers remains an open question. Still, this let us motivate the multi-task approach that we study in Chapter 4.

Chapter 4

Multi-task setting

The models studied in Chapter 3 are only single-task models. Effectively, the national setting let us introduce a prototype of a load forecasting model and, at the 4 lower aggregation levels of Section 2.5.2, we only replicated the national model for the multivariate forecasting problems to obtain a collection of single-task models with heterogeneous inputs and adapted hyperparameters.

To estimate the coefficients of the different single-task models at a given aggregation level, we solve Problem 3.25. The objective and the regularization are separable since they are written as a sum over the different tasks and consequently, every learning problem in Chapter 3 for a given area is isolated, or independent, from the others.

In Section 3.5.5 and Section 3.5.6, we have had the opportunity to emphasize the similarities between the coefficients learned for the different substations. Even though it was definitely expected, the model found by itself resembling relationships between the load and the inputs at the different substations. The main question addressed in this chapter is the following :

> **Can we leverage this similarity structure in the multi-task setting in order to guide the learning with constraints and regularization ?**

Put differently, this question asks whether information sharing between the different areas within a given aggregation level can help to produce better forecasts.

By coupling the different tasks, we aim at reducing the complexity of the multi-task model in some directions and avoid overfitting while simultaneously increasing the complexity in other directions, altogether to obtain a better generalization performance. It is possible for instance to reduce the complexity by mutualizing some parameters between the substations or by regularizing their differences. On the contrary, we may increase the complexity in other directions by reducing the regularization hyperparameters. Sharing information between the substations may also reduce the necessary amount of data in order to avoid dataset shift. Therefore, we study in this chapter the possibility of an actual multi-task model, that is to say a model that is not a mere collection of independent single-task models.

In Section 4.1, we go further in the analysis of the similarity structure among the single-task models computed in Section 3.5.5 and Section 3.5.6. Related work and off-the-shelf benchmarks for multi-task problems are presented in Section 4.2.

In Section 4.3, we formally present possible ways to couple the models in a multi-task setting. In Section 4.4 and Section 4.5, we present multi-task models that rely on a structural assumption on the coefficient matrix, namely that some of its columns are constrained to be identical or that they all lie in a low-dimensional subspace. Finally in Section 4.6, we consider a multi-task model whose difference with the models of Chapter 3 resides in the loss function.

4.1 Structure of the independent models

We compared in Section 2.4.2 the distribution of the input data for different substations. In Section 3.5.5 and Section 3.5.6, we illustrated the distribution of the coefficients learned by independent single-task models at the local level. In this section, we additionally motivate the multi-task setting by presenting the similarities in the forecasts and in the residuals of these models. The graphs presented in this section follow the estimation of the models with the training years 2013 to 2015 and the test year 2016.

4.1.1 Similarities between models

Because we are interested in coupling the models of the different substations, we first analyze the similarities between the estimated coefficients.

Clustering A natural illustration of the potential similarities between the models is to cluster the learned coefficient vectors and look for a visually apparent structure. However, we do not have a clear interpretation of the clusters presented in Figure F.68. We study in more details the possibility of clustering the coefficients in Section 4.4.

Rank of the coefficient matrix Alternatively, we propose in Figure 4.1 an illustration of the spectrum of the learned coefficient matrix :

$$\boldsymbol{B} := \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{C} \end{bmatrix} \in \mathbb{R}^{p,K}, \tag{4.1}$$

where $\boldsymbol{A}$ is the bloc corresponding to the features shared by all the substations and $\boldsymbol{C}$ is the bloc corresponding to the individual features in the model defined by Equation (3.25), with the parametrization for the substations level described in Section 3.4. Additionally, the spectrum of the prediction matrices are illustrated in Figure F.69.

In both cases, a significant part of the spectrum is localized in the first components of the approximations. This leads us to question whether these first components could be shared by the substations. We develop this idea with a low-rank constraint on the coefficient matrix in Section 4.5.

4.1.2 Commonly structured errors

While Section 4.1.1 is dedicated to the similarities between the local models, we study in this section the structure of their residuals.

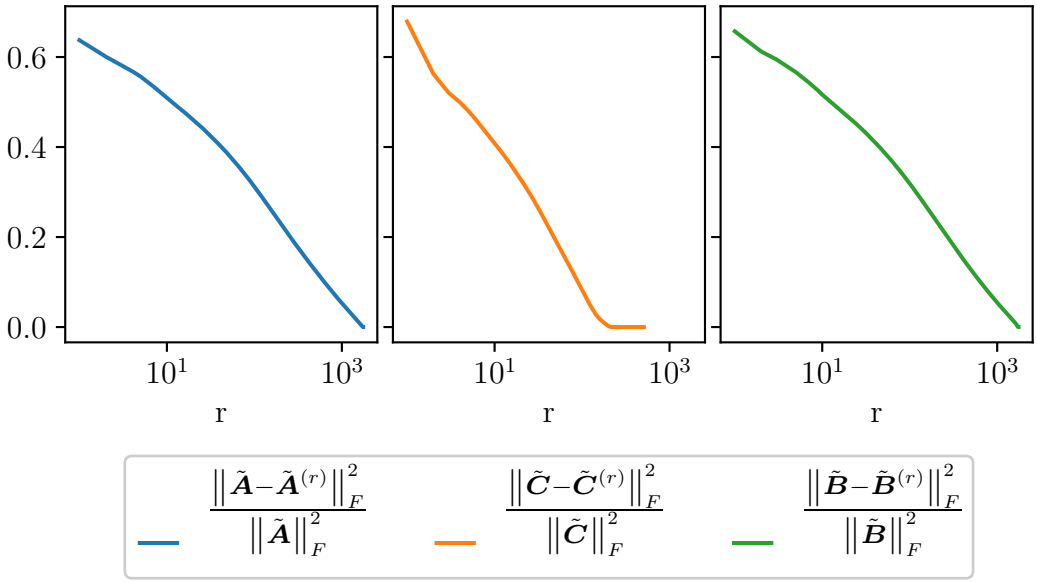


FIGURE 4.1: Singular values of the coefficient matrices
Norm of the residuals after subtracting the best rank-r approximation of the coefficient matrices whose rows have been centered. Given $M \in \{\boldsymbol{B}, \boldsymbol{A}, \boldsymbol{C}\}$, we denote $\tilde{M}$ the same matrix after its rows have been centered and for $r \in \mathbb{N}$, the matrix $\tilde{M}^{(r)}$ is the closest rank-r approximation of $\tilde{M}$ in terms of the Frobenius norm (*c.f.* Figure 2.22 for supplementary details). With the parametrization of the local models described in Chapter 3, the matrix $\boldsymbol{A}$ has dimension (3379, 1751), the matrix $\boldsymbol{C}$ containing the coefficients related to the past loads and the weather conditions has dimension (510, 1751) and the concatenation $\boldsymbol{B}$ has dimension (3889, 1751). The three functions equal 1 for $r = 0$ but the curves begin at $r = 1$ because of the logarithmic scale of the x-axis.

Spatial Correlation A natural way ro represent the residual correlations between the different substations is to represent them on maps. It seems indeed reasonable to assume that there is a higher probability to be correlated for nearby substations.

We present these correlations for 9 different substations in Figure 4.2. A map is associated with one substation $k \in [\![1, K]\!]$, indicated by the black circle, and the color in another area ℓ represents the empirical correlations in the residuals with the corresponding substation ℓ :

$$\mathrm{corr}_{\mathrm{test}}(\mathsf{y}_k - \hat{\mathsf{y}}_k, \mathsf{y}_\ell - \hat{\mathsf{y}}_\ell) := \frac{\mathbb{E}_{\mathrm{test}}[(\mathsf{y}_k - \hat{\mathsf{y}}_k)(\mathsf{y}_\ell - \hat{\mathsf{y}}_\ell)]}{\sqrt{\mathbb{V}\mathrm{ar}_{\mathrm{test}}[\mathsf{y}_k - \hat{\mathsf{y}}_k]\mathbb{V}\mathrm{ar}_{\mathrm{test}}[\mathsf{y}_\ell - \hat{\mathsf{y}}_\ell]}}, \tag{4.2}$$

where $\mathbb{V}\mathrm{ar}$ denotes the variance, y_k, $\mathsf{y}_\ell \in \mathbb{R}$ are the normalized loads defined in Equation (3.33) and $\hat{\mathsf{y}}_k$, $\hat{\mathsf{y}}_\ell \in \mathbb{R}$ are their estimates.

For a given substation, the closest substations in Figure 4.2 are not necessarily the only ones that present an important positive correlation. However, a small geographical distance between two substations seems to be a strong indicator of potential correlations. Since we are interested in this section in sharing information between different models for a multi-task models, it is essential to ask which substations should be coupled.

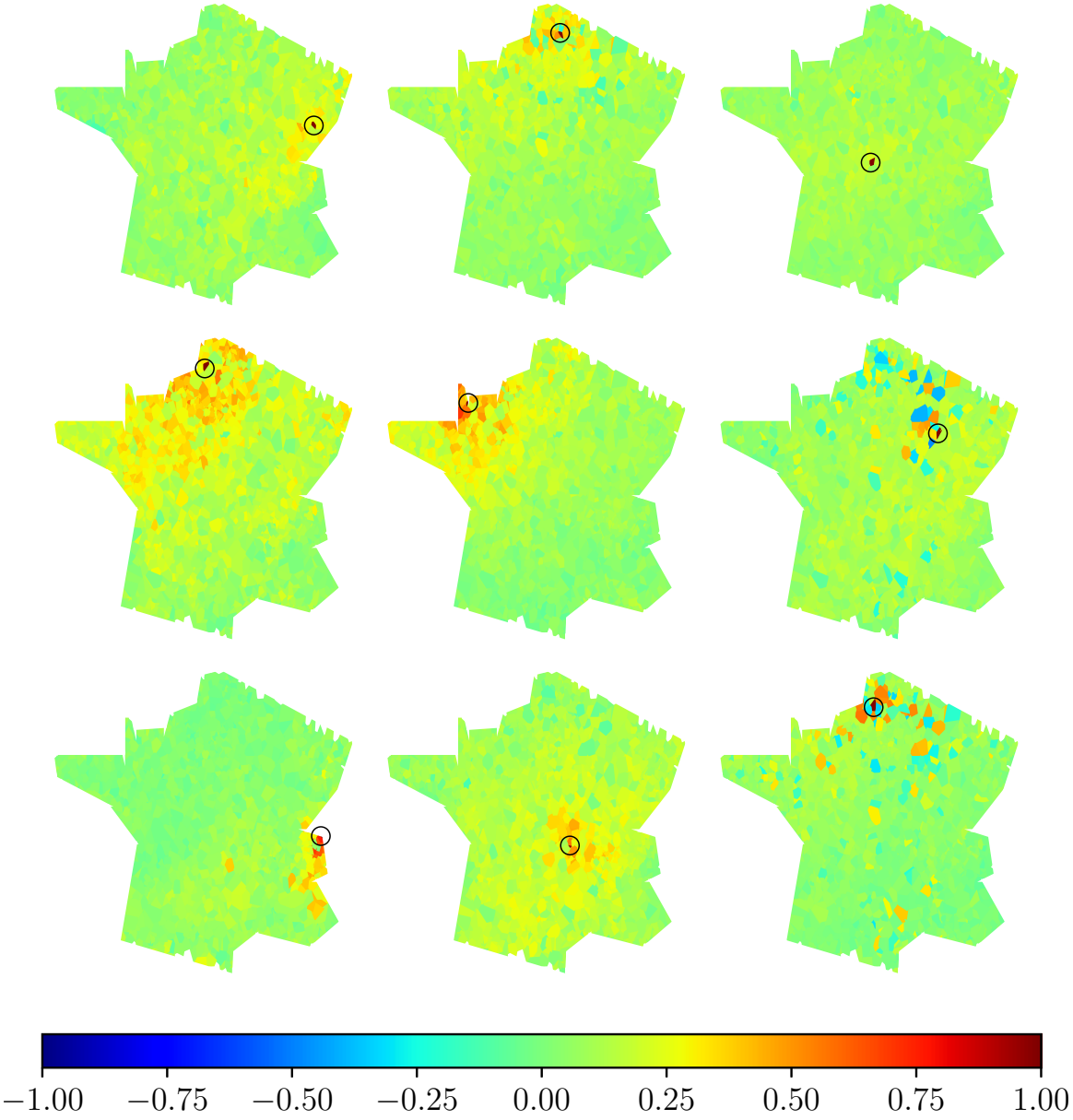


FIGURE 4.2: Spatially correlated residuals
Out of honesty, we illustrate the correlation of the residuals in the test year 2016 for the 9 first substations in the database ordered alphabetically. The residuals of the substation in the center are positively correlated with the residuals of geographically close substations. On the contrary, there are negative correlations for the substation in the center right. As for the substation in the top right-hand corner, it has little correlation with the others.

To emphasize the important correlations between the substations and their neighbors, let $N^k_{\nu'}$ denote the ν'-th geographically closest neighbor of a substation k. The average over the substations of the correlation between a substation k and its ν'-th

closest neighbor :

$$\rho_{\nu'} := \frac{1}{K} \sum_{k=1}^{K} \mathrm{corr}_{\mathrm{test}}(\mathsf{y}_k - \hat{\mathsf{y}}_k, \mathsf{y}_{N^k_{\nu'}} - \hat{\mathsf{y}}_{N^k_{\nu'}}), \tag{4.3}$$

is illustrated in Figure 4.3, as well as the average correlation with the ν closest neighbors :

$$\tau_\nu := \frac{1}{\nu} \sum_{\nu'=1}^{\nu} \rho_{\nu'}. \tag{4.4}$$

Figure 4.3 confirms that positive correlations decrease on average with the geographical distance separating two substations.

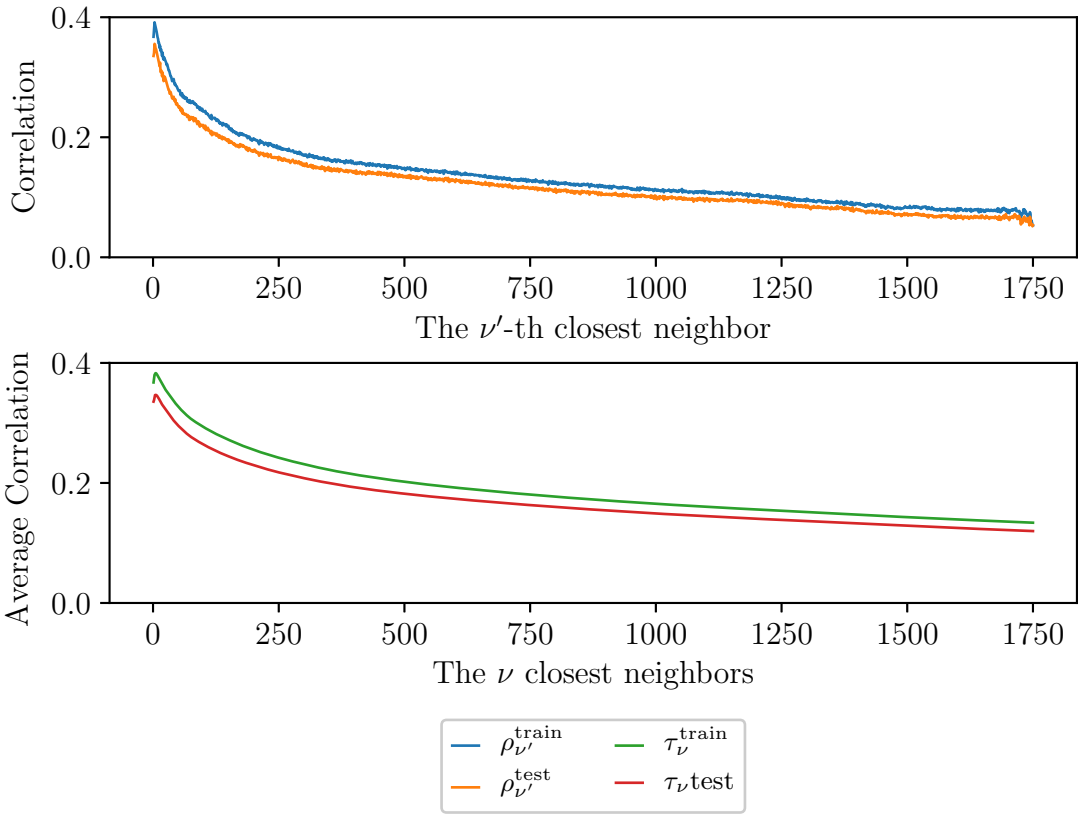


FIGURE 4.3: Residual correlations with closest neighbors

(*top*) Average over the substations of the residual correlation in the training and the test sets with the ν'-th geographically closest neighbor.

(*bottom*) Average over the substations of the residual correlation in the training and the test sets with the ν geographically closest neighbors.

Temporal correlation In addition to the spatial correlations, we illustrate in Figure 4.4 the average quantiles of the residuals and their norms for different values of the hour of the week and over the test year 2016. The night and the summer

are better predicted than the daytime and the winter for all the quantiles. The fact that overestimation and underestimation of the loads occur simultaneously for the 5 quantiles indicates the presence of commonly structured errors : the hardest moments to predict seem common to a majority of the substations. We explore this idea in Section 4.6.

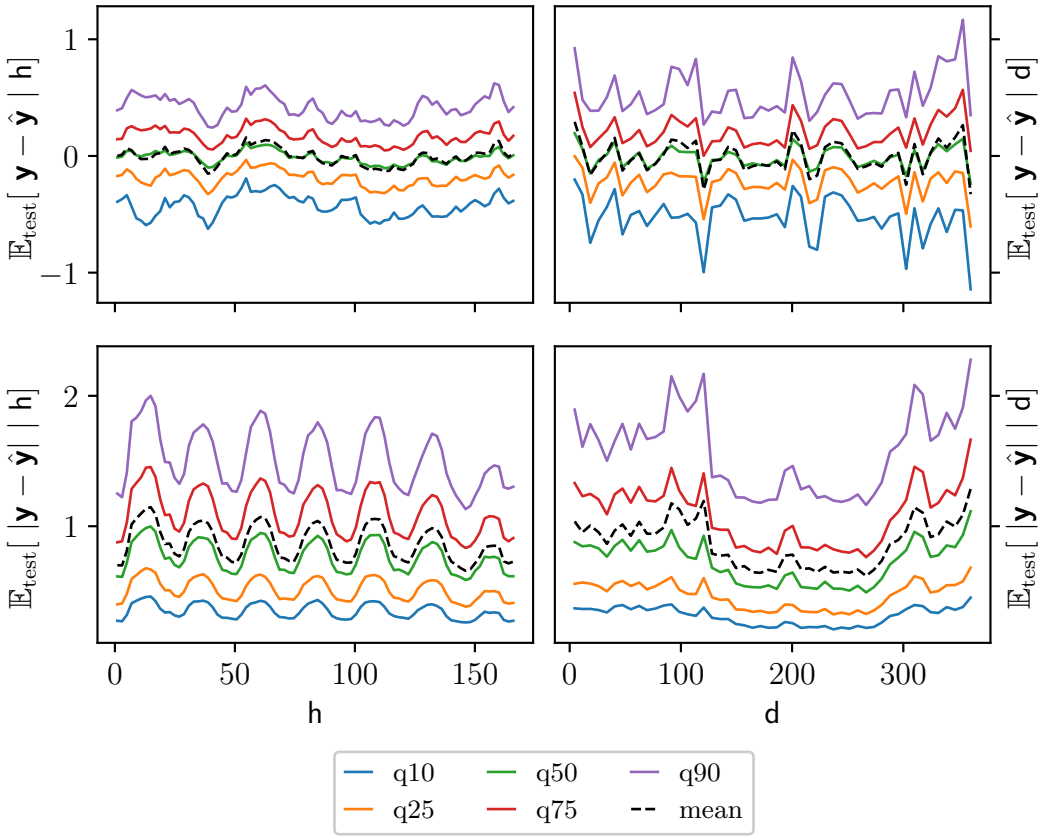


FIGURE 4.4: Temporally structured residuals
Quantiles and mean over the substations of the average residuals and average norm of the residuals conditioned on the different values of the hour of the week h and of the day of the year d over the test year 2016.

(*top left*) Average residuals conditioned with the hour of the week $\mathbb{E}_{\text{test}}[\,\mathbf{y}-\hat{\mathbf{y}}\mid \mathsf{h}]$.

(*bottom left*) Average norm of the residuals conditioned with the hour of the week $\mathbb{E}_{\text{test}}[\,|\mathbf{y}-\hat{\mathbf{y}}|\mid \mathsf{h}]$.

(*top right*) Average residuals conditioned with the day of the year $\mathbb{E}_{\text{test}}[\,\mathbf{y}-\hat{\mathbf{y}}\mid \mathsf{d}]$.

(*bottom left*) Average norm of the residuals conditioned with the day of the year $\mathbb{E}_{\text{test}}[\,|\mathbf{y}-\hat{\mathbf{y}}|\mid \mathsf{d}]$.

The hour of the week $\mathsf{h}=0$ corresponds to Monday at 00:01 and the day $\mathsf{d}=0$ is the 1st of January. The norms of the residuals are smaller during the 7 nights of the week and from $\mathsf{d}=120$, that is the end of April, to $\mathsf{d}=270$, the end of September. The goal of these graphs is to show that the quantiles have similar variations, which means that the most difficult periods of the week and of the year are the same for all the substations.

4.2 Related work - Multi-task learning

Before the advent of the Machine Learning algorithms, analyzing the characteristics of multiple individuals (or tasks) in a population has been the object of interest of various statistical methods. In order to model a structured population, for instance partitioned into different groups, the multilevel models were proposed to estimate the relationships between those characteristics while taking into account the multilevel structure. This led to the development of statistical frameworks including the mixed effect models, a combination of fixed and random effect models and in particular with the Bayesian approach, to Bayesian hierarchical models [Good, 1980]. Given that many of the ideas employed nowadays in multi-task learning originate from these statistical models, we first propose a description of the main hierarchical models encountered in the literature since the 1950s. Thereby, we extend the presentation of Section 2.9.4 about the Generalized Additive and Linear Models.

Mixed effect models Consider K tasks and let y_k denote the scalar target variable and $\boldsymbol{\xi}^{(k)}$ a D-dimensional vector of inputs for task $k \in [\![1, K]\!]$. In a general Linear Model (**gLM**), or multivariate-output regression model, the k-th component y_k of the target vector $(\mathsf{y}_1, \ldots, \mathsf{y}_K)$ follows the model :

$$\mathsf{y}_k = \langle \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta}^{(k)} \rangle + \varepsilon_k, \tag{4.5}$$

where ε_k contains the error and $\boldsymbol{\beta}^{(k)}$ is an unknown fixed (non-random) coefficient vector to be estimated from the data, typically via the minimization of a penalized error. Besides, a gLM assumes that the vector $(\varepsilon_1, \ldots, \varepsilon_K)$ follows a normal distribution.

The extension to Generalized Linear Models (**GLM**) is the same as in the single-task setting described in Section 2.9.4 : the target vector $(\mathsf{y}_1, \ldots, \mathsf{y}_K)$ is assumed to be generated from an overdispersed exponential family and its mean value is related to the covariates via :

$$\mathbb{E}[\mathsf{y}_1, \ldots, \mathsf{y}_K] = g^{-1}\left(\langle \boldsymbol{\xi}^{(1)}, \boldsymbol{\beta}^{(1)} \rangle, \ldots, \langle \boldsymbol{\xi}^{(K)}, \boldsymbol{\beta}^{(K)} \rangle\right), \tag{4.6}$$

where g is the link function like in Equation (2.31). Note that it is sometimes required in gLM and GLM that the inputs are identical for all the tasks. Otherwise, the inputs can be of the form :

$$\boldsymbol{\xi}^{(k)} := \begin{bmatrix} \boldsymbol{\xi}^{(0)} \\ \boldsymbol{\zeta}^{(k)} \end{bmatrix} \in \mathbb{R}^D, \tag{4.7}$$

where $\boldsymbol{\xi}^{(0)}$ is *common* to all the tasks and $\boldsymbol{\zeta}^{(k)}$ are *task-specific* inputs.

In the rest of this section, like in most papers about mixed-effect models, we assume for simplicity that the link function is the identity. The extension to Generalized Additive Models (**GAM**) is also similar to the case of univariate models, and we emphasize the distinction between the linear and the nonlinear effects :

$$\mathsf{y}_k = \langle \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta}^{(k)} \rangle + \sum_{d=1}^{D} f_d^{(k)}(\xi_d^{(k)}) + \varepsilon_k, \tag{4.8}$$

where the $f_d^{(k)}$ are unknown fixed (non-random) functions to be estimated, possibly non-parametric. Similarly to the inputs in Equation (4.7), the effects may be decomposed into *local* and *common* parts :

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\alpha}^{(0)} + \boldsymbol{\gamma}^{(k)} \qquad f_d^{(k)} = f_d^{(0)} + g_d^{(k)}, \tag{4.9}$$

to account for effects *common* to all the tasks and for *task-specific* effects.

So far, all the parameters of the model are fixed quantities to be estimated from the data. That is why Fisher [1919] introduced Random Effects Models, to take into account the correlations between the different tasks and deal with overdispersed data. Combined with fixed effects, they are called Mixed Effects Models and have become subsequently a major branch of Statistics[1]. The mixed effects modeling version of the gLM given in Equation (4.5) is :

$$\mathsf{y}_k = \langle \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta}^{(k)} \rangle + \langle \boldsymbol{\chi}^{(k)}, \boldsymbol{\delta}^{(k)} \rangle + \varepsilon_k, \tag{4.10}$$

where ε_k contains the error, $\boldsymbol{\xi}^{(k)}$ and $\boldsymbol{\chi}^{(k)}$ are two known vectors of inputs for task k, $\boldsymbol{\beta}^{(k)}$ is a *fixed* vector of coefficients to be estimated from the data and $\boldsymbol{\delta}^{(k)}$ is the *random* coefficient vector for task k. Furthermore, $(\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(K)})$ is assumed to follow a *prior* distribution parametrized by a hyperparameter $\boldsymbol{\theta}$.

The mixed-effects version of the GLM in Equation (4.6) leads to the so-called Generalized Linear Mixed Model [Breslow and Clayton, 1993] and the extension of the GAM in Equation (4.8) gives rise to the Generalized Additive Mixed Models, for example with a Wiener process prior on the non-linear effects [Lin and Zhang, 1999]. We also denote the parameter of the prior $\boldsymbol{\theta}$ for simplicity.

Depending on the structure of the modeled population and in particular its hierarchical organization, the hyperparameter $\boldsymbol{\theta}$ can be random and the parameters of the *hyperprior* distribution of $\boldsymbol{\theta}$ can themselves be parametrized with a *hyperhyperprior*, which is either fixed before any data is observed, random with an additional layer of hyperparameters, or to be estimated from the data via its most likely value. The latter approach is called the empirical Bayes method, also known as maximum marginal likelihood, and corresponds to an approximation of a fully Bayesian setting [Robbins, 1956].

A complementary type of analysis of structured data lead to the development of the Hierarchical Linear Models (**HLM**) [Lindley and Smith, 1972; Raudenbush and Bryk, 2002], a subclass of Hierarchical Bayesian Network [Pearl, 1985], particularly appropriate for the modeling of data organized in nested groups with a specific attention devoted to the variance of the estimated coefficients.

Complex objects require complex modeling which justify the use of these second (prior) and third (hyperprior) order probabilities. Following Occam's razor principle and quoting Good [1980], one should "stop when the guessed expected utility of going further becomes negative if the cost is taken into account".

The Multi-task paradigm Multi-task learning emerged 20 years ago [Baxter, 2000; Caruana, 1997], from the ideas developed in relation with Mixed Effects Models, as a branch of Machine Learning focused on the estimation of several tasks in

[1]Note that the terminology for fixed and random effects may vary between statistical models [Gelman et al., 2005]. We use the adjective *fixed* to designate a non-random quantity.

parallel, with the objective of finding an inductive bias, or alternatively learning to learn [Heskes, 2000], by leveraging a common structure shared by the multiple tasks. It is different from Transfer Learning that considers the more general ambition of transferring knowledge between tasks, possibly sequentially from some source tasks already learned to a target task yet to be learned.

With the multi-task approach, different problems are assumed to be related, meaning that they share a common underlying structure. They are coupled to share the information gathered for each individual task and improve the generalization performance of all of them by resisting the fact that data may be scarce in some regions of the input spaces. Indeed, as the complexity of the model increases in a single-task problem to reduce the *approximation* error, the data relevant to estimate each parameter becomes scarcer and the risk of overfitting is more important : the *estimation* error may become larger. It is the case for example for the univariate features introduced in Section 3.1.1 when the number of knots becomes too large or for the interactions that we introduced in Section 3.1.2. The bet made by multi-task models is that sharing information between different tasks can potentially prevent overfitting in these more complex models and help to estimate robust coefficients.

The questions to be addressed when considering a multi-task problem are :

> **What to expect from a coupling ? Which tasks are related ? What is shared by the different tasks ? How to couple them ?**

Early works [Baxter, 2000; Ben-David and Schuller, 2003; Maurer, 2009] tried and provided theoretical bounds to answer the first question. Although crucial, few generic methods exist to determine which tasks are related and which ones are outlier tasks. We motivated the multi-task setting in Section 4.1 but mostly rely on empirical results to answer this second question. Indeed, imposing a common structure on two unrelated tasks can lead to *negative transfer* [Perkins et al., 1992], that is a deterioration of the performances on both tasks.

Zhang and Yang [2017] propose a survey of the larger literature concerning common structures that can be shared between tasks and possible ways to proceed. Implicitly, available methods rely on different assumptions about this underlying common structure. Rai et al. [2012] distinguish two possible structures : they propose to denote *task structure* the similarities encountered in the models learned for different tasks, without taking into account the distribution of the inputs and outputs, and to denote *output structure* similarities in the residuals produced by the models corresponding to the different tasks. Besides, these structural assumptions may be formulated within a Bayesian framework, in connection with the early works on multivariate-output regression, or may be integrated in the learning problem without any probabilistic considerations.

Task structure Essentially, a common task structure corresponds to the presence of similarities in the parameters of the estimation models used for the different tasks. These assumptions are enforced by leveraging a structural regularization, constraints or common feature representations.

First, the parameters can be close in a geometric sense, in which case a clustering of the coefficients is relevant [Bakker and Heskes, 2003; Evgeniou et al., 2005; Evgeniou and Pontil, 2004]. Alternate minimization procedures have been proposed as

well as convex relaxations of the resulting optimization problems [Jacob et al., 2009]. For each cluster, or groups of coefficient vectors, the data that can be leveraged to estimate the average coefficients in a given cluster extends to the data available for all the tasks in this cluster. Note that such an assumption often ignores the negative correlations that can exist between different tasks.

Secondly, an alternative assumption considers that the coefficient vectors of different tasks span a low-dimensional subspace, are close to a low-dimensional subspace [Ando and Zhang, 2005] or to low-dimensional subspaces [Kumar and Daume III, 2012]. Convex relaxation of optimization problems with low-rank constraints have been considered, notably with the introduction of the trace-norm in the resulting optimization problems [Pong et al., 2010], leading to matrices of coefficients with a sparse spectrum. This constraint was also considered within Bayesian frameworks. For instance, the subspace clustering problem considered by Elhamifar and Vidal [2013] was equally considered by Wang et al. [2015] with a Bayesian formulation and [Wipf, 2014] adopted a Bayesian formulation and a variational approximation to assess the possibility of smoothing local minimizers for matrix rank minimization.

Thirdly, the common structure of the parameters for different tasks can reside in their nonzero components. Group-Lasso penalizations have been considered for joint variable selection [Obozinski et al., 2010] based on the generalized group version of LASSO [Bakin et al., 1999; Yuan and Lin, 2006] or to selectively screen which variables should be part of the shared components [Ando and Zhang, 2005], thereby partially addressing the problem of *negative transfer*.

Output structure The second structure identified by Rai et al. [2012] refers to similarities in the residuals of different tasks when conditioning on the inputs. This can certainly be due to an insufficient amount of information in the inputs, a limited expressiveness of the models present in the hypothesis space, or to a bad choice of regularization, but it can also result from naturally correlated noises in the different tasks. In this case, it is relevant to model the covariance structure in the outputs conditioned on the inputs by introducing a structured loss function or a structured regularization [Rai et al., 2012; Rothman et al., 2010].

Note that we have limited the presentation to shallow multi-task problems while a vast literature about multi-task neural networks has been written in the last decade. Multivariate-output tree-based models were also considered [Dumont et al., 2009]. However, for the specific problem of load forecasting, our attempts at finding a deep or tree-based multi-task benchmark did not result in better performances than the independent models of Chapter 3, which we consequently keep as a baseline in this chapter to assess the relevance of the multi-task approach.

4.3 Framework for multi-task learning

In this section, we introduce a general framework used thereafter to present the multi-task learning problems that we consider. The presentation is restricted to settings in which learning problems are formulated as optimization problems.

Notations We consider a problem with K tasks. We denote $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^{D_0}$ a set of inputs common to all the tasks (*e.g.* the hour of the week) and for each task $k \in [\![1, K]\!]$, we denote $\boldsymbol{\zeta}^{(k)} \in \mathbb{R}^{D-D_0}$ a set of individual inputs (*e.g.* the past loads or the temperatures at a given weather station), where $D \in \mathbb{N}^*$, $D_0 \in \mathbb{N}$ and $D \geq D_0$. The vector $\boldsymbol{\xi}^{(k)} \in \mathbb{R}^D$ denotes the concatenation of $\boldsymbol{\xi}^{(0)}$ with $\boldsymbol{\zeta}^{(k)}$:

$$\boldsymbol{\xi}^{(k)} := \begin{bmatrix} \boldsymbol{\xi}^{(0)} \\ \boldsymbol{\zeta}^{(k)} \end{bmatrix} \in \mathbb{R}^D. \tag{4.11}$$

Following the notations of Section 3.3, we denote $\mathbf{x}^{(0)} \in \mathbb{R}^{p_0}$ the vector of $p_0 \in \mathbb{N}$ *common* covariates obtained with the *common* inputs $\boldsymbol{\xi}^{(0)}$ and the feature engineering of Section 3.1. For each task $k \in [\![1, K]\!]$, the vector of *local* covariates built from $\boldsymbol{\zeta}^{(k)}$ is denoted $\mathbf{z}^{(k)} \in \mathbb{R}^{p-p_0}$. Similarly to Equation (4.11), the vector $\mathbf{x}^{(k)} \in \mathbb{R}^p$ denotes the concatenation of $\mathbf{x}^{(0)}$ with $\mathbf{z}^{(k)}$:

$$\mathbf{x}^{(k)} := \begin{bmatrix} \mathbf{x}^{(0)} \\ \mathbf{z}^{(k)} \end{bmatrix} \in \mathbb{R}^p. \tag{4.12}$$

We also define for each task $k \in [\![1, K]\!]$, the coefficient vectors $\boldsymbol{a}^{(k)} \in \mathbb{R}^{p_0}$ and $\boldsymbol{c}^{(k)} \in \mathbb{R}^{p-p_0}$ respectively for the *common* and the *local* covariates like in Equation (3.25), $\boldsymbol{b}^{(k)} \in \mathbb{R}^p$ as the concatenation of $\boldsymbol{a}^{(k)}$ and $\boldsymbol{c}^{(k)}$ and the corresponding matrices of coefficients as :

$$\boldsymbol{A} := \left[\boldsymbol{a}^{(1)} \dots \boldsymbol{a}^{(K)}\right] \in \mathbb{R}^{p_0,K}, \tag{4.13}$$

$$\boldsymbol{C} := \left[\boldsymbol{c}^{(1)} \dots \boldsymbol{c}^{(K)}\right] \in \mathbb{R}^{p-p_0,K}, \tag{4.14}$$

$$\boldsymbol{B} := \left[\boldsymbol{b}^{(1)} \dots \boldsymbol{b}^{(K)}\right] := \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{C} \end{bmatrix} \in \mathbb{R}^{p,K}. \tag{4.15}$$

For each task k, the target variable $\mathsf{y}_k \in \mathbb{R}$ is modeled with a function parametrized by $\boldsymbol{b}^{(k)}$:

$$f_{\boldsymbol{b}^{(k)}} : \mathbb{R}^D \to \mathbb{R}, \tag{4.16}$$

and given $\boldsymbol{\Xi} := (\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(K)}) \in \mathbb{R}^{D,K}$, we denote the target vector and its multivariate estimate :

$$\mathbf{y} := \begin{bmatrix} \mathsf{y}_1 \\ \vdots \\ \mathsf{y}_K \end{bmatrix} \in \mathbb{R}^K \quad \text{and} \quad F_{\boldsymbol{B}}(\boldsymbol{\Xi}) := \begin{bmatrix} f_{\boldsymbol{b}^{(1)}}(\boldsymbol{\xi}^{(1)}) \\ \vdots \\ f_{\boldsymbol{b}^{(K)}}(\boldsymbol{\xi}^{(K)}) \end{bmatrix} = \begin{bmatrix} \langle \mathbf{x}^{(1)}, \boldsymbol{b}^{(1)} \rangle \\ \vdots \\ \langle \mathbf{x}^{(K)}, \boldsymbol{b}^{(K)} \rangle \end{bmatrix} \in \mathbb{R}^K. \tag{4.17}$$

Non-separable Model The most general form of multi-task model that we consider is :

$$\mathcal{M} \quad \sim \quad F_{\boldsymbol{B}}(\boldsymbol{\Xi}) \quad \in \quad \mathbb{R}^K, \tag{4.18}$$

and for the minimization of a regularized empirical risk, the most general optimization problem that we consider is defined for a set of observations $(\boldsymbol{y}_i)_{i=1,\dots,n} \in \mathbb{R}^{n,K}$ and $(\boldsymbol{\Xi}_i)_{i=1,\dots,n} \in \mathbb{R}^{n,D,K}$ as :

$$\begin{aligned} \min_{\boldsymbol{B} \in \mathbb{R}^{p,K}} \quad & \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(\boldsymbol{y}_i, F_{\boldsymbol{B}}(\boldsymbol{\Xi}_i)\right) + \Omega(\boldsymbol{B}), \\ \text{s.t.} \quad & \boldsymbol{B} \in \boldsymbol{\Theta}, \end{aligned} \tag{4.19}$$

for some general

$$\text{loss function} \quad \mathcal{L} : \mathbb{R}^{K,2} \to \mathbb{R}_+, \tag{4.20}$$

$$\text{regularization} \quad \Omega : \mathbb{R}^{p,K} \to \mathbb{R}_+, \tag{4.21}$$

$$\text{and set of constraints} \quad \boldsymbol{\Theta} \subset \mathbb{R}^{p,K}. \tag{4.22}$$

The coupling between the different tasks in an optimization problem is determined by a set of restrictions on the form of the model $(f_{\boldsymbol{b}^{(k)}})_{k=1,\dots,K}$, the loss $\mathcal{L}$, the regularization Ω and the set of constraints $\boldsymbol{\Theta}$. In particular, problems with less coupling are defined with particular cases of Problem 4.19. Without explicitly mentioning it, we already saw in Chapter 2 and Chapter 3 two restrictions in optimization problems that prohibit the coupling between the different tasks, they are formalized below.

Separable over the substations The independent models defined in Chapter 3 can be completely decomposed into subproblems like Problem 3.20 since they have :

- a separable loss for the load vector $\mathbf{y} \in \mathbb{R}^K$ and its estimate $\hat{\mathbf{y}} \in \mathbb{R}^K$:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} \ell(\mathsf{y}_k, \hat{\mathsf{y}}_k), \tag{4.23}$$

  where $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ is the same loss for the different tasks. Indeed, given $n \in \mathbb{N}^*$, a matrix of target observations $\boldsymbol{Y} \in \mathbb{R}^{n,K}$ and design matrices $\boldsymbol{\mathcal{X}} := (\boldsymbol{X}^{(k)})_{k=1,\dots,K} \in \mathbb{R}^{n,p,K}$, we considered for Problem (3.25) the empirical risk :

$$\frac{1}{2n} \sum_{k=1}^{K} \left\| \boldsymbol{y}^{(k)} - \boldsymbol{X}^{(k)} \boldsymbol{b}^{(k)} \right\|_2^2. \tag{4.24}$$

- a separable regularization :

$$\Omega(\boldsymbol{B}) = \sum_{k=1}^{K} \omega(\boldsymbol{b}^{(k)}), \tag{4.25}$$

  where $\omega : \mathbb{R}^p \to \mathbb{R}_+$ is the same regularizer for the different tasks.

- and a separable set of constraints :

$$\boldsymbol{\Theta} = \prod_{k=1}^{K} \Theta, \tag{4.26}$$

  where $\Theta \subset \mathbb{R}^p$ is identical for the different tasks.

Separation of the substations and partition of the input domain For the GAM introduced in Section 2.9.4, Goude et al. [2013] built one model per hour[2]

[2] More exactly, their time series are recorded every 10 minutes instead of every hour and they use the hour of the day and the day of the week instead of the hour of the week but this does not modify the scope of this paragraph. They have consequently $24 \times 6 = 144$ submodels.

and per substation. Thouvenot [2015] proceeded similarly. However, they do not consider other interactions between input variables. The hour $\mathsf{h} \in H := [\![0, 167]\!]$, which is one of the inputs, is used to divide the model into a collection of independent submodels with distinct parameters $(\boldsymbol{B}_h)_{h\in H} \in \mathbb{R}^{p,K,|H|}$. Indeed, the problem that they consider is unconstrained and we can write the model and the regularization as :

$$\mathcal{M}_1 \quad \sim \quad \sum_{h\in H} 1_{\mathsf{h}=h} F_{\boldsymbol{B}_h}(\boldsymbol{\Xi}) \quad \in \quad \mathbb{R}^K, \tag{4.27}$$

$$\Omega((\boldsymbol{b}_h^{(k)})_{k=1,\dots,K,h\in H}) \quad = \quad \sum_{k=1}^{K}\sum_{h\in H} \omega(\boldsymbol{b}_h^{(k)}), \tag{4.28}$$

where $\omega : \mathbb{R}^p \to \mathbb{R}_+$. This decomposition into subproblems amounts to partitioning the database into different sets of samples corresponding to the different hours.

Scope of Chapter 4 In this chapter, we are interested in multi-task problems without the restrictions identified in Equations (4.23), (4.25), (4.26), (4.27) and (4.28). This aims at coupling the substations while it was not permitted in Chapter 3.

Non-separable sets of constraints are considered in Section 4.4.1 and Section 4.5.1. Non-separable regularizations are presented in Section 4.4.2 and Section 4.5.2. Finally, a problem with a non-separable loss is introduced in Section 4.6.

For the national level, the RTE regions, the Administrative regions and the districts level, the parametrization of the models in this chapter is the same as in Chapter 3, it is described in Tables 3.2, 3.3 and 3.4 . The number of knots and the interaction of the load forecasting models for the substations are also identical to Chapter 3 and given in Tables F.2, F.3 and F.4. However, we precise in this chapter when the set of regularization hyperparameters may change.

4.4 Clustering of the substations

This section is dedicated to studying the possibility of clustering the substations such that in each group, the coefficient vectors used to forecast the normalized loads are close, where the normalization is given in Equation (3.33). In Section 4.4.1, we impose a strict equality condition : the same model is learned for all the tasks in a same cluster. A possible relaxation of this hard constraint is proposed in Section 4.4.2. Finally, experiments are presented in Section 4.4.3.

The models and results of this section are extracted from the internal report written by Duchemin [2018], after our collaboration during his internship in our team, at École Nationale des Ponts et Chaussées. For pragmatic and confidentiality reasons, the database used for the experiments in Section 4.4.3 is different from the database provided by RTE. Instead, the experiments were performed with the GEFCom 2012 database [Hong et al., 2014].

This database was produced for a competition and little information is given about the substations and the weather stations. As a consequence, selecting the relevant weather stations to forecast each load was one of the expected challenges and the models described in this section precisely perform variable selection.

4.4.1 Formulation of the hard clustering problem

Clustering For the hard clustering problem, we consider the empirical risk given in Equation (4.24) and we make the assumption, similarly to [Bakker and Heskes, 2003], that the different tasks can be grouped into $R \in \mathbb{N}^*$ clusters such that all the tasks in one cluster are forecast with an identical coefficient vector. The set of constraints that we consider has consequently the non-separable form :

$$\boldsymbol{\Theta}_{\text{HC}} := \left\{ \boldsymbol{B} \in \mathbb{R}^{p,K} \quad | \quad \left| \left\{ \boldsymbol{b}^{(1)}, \dots, \boldsymbol{b}^{(K)} \right\} \right| \leq R \right\}, \tag{4.29}$$

where $\left| \left\{ \boldsymbol{b}^{(1)}, \dots, \boldsymbol{b}^{(K)} \right\} \right|$ denotes the cardinal of the set of column vectors $\left\{ \boldsymbol{b}^{(1)}, \dots, \boldsymbol{b}^{(K)} \right\}$ and typically, $R \ll K$.

Given $\boldsymbol{B} \in \boldsymbol{\Theta}_{\text{HC}}$, let $\boldsymbol{U} \in \mathbb{R}^{p,R}$ whose columns are the R possible values for the columns of $\boldsymbol{B}$. Also, let $\boldsymbol{V} \in \{0,1\}^{K,R}$ such that for any $k \in [\![1, K]\!]$, if the k-th column of $\boldsymbol{B}$ is $\boldsymbol{u}^{(r)}$ with $r \in [\![1, R]\!]$, then the k-th row of $\boldsymbol{V}$ is the indicator vector $\boldsymbol{v}_k \in \{0,1\}^R$ with a non-zero coefficient in position r. Thereby, we can write $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T$. In other words, an explicit parametrization of $\boldsymbol{\Theta}_{\text{HC}}$ with $pR + K$ degrees of freedom is given by :

$$\boldsymbol{\Theta}_{\text{HC}} = \left\{ \boldsymbol{U}\boldsymbol{V}^T \quad | \quad \boldsymbol{U} \in \mathbb{R}^{p,R}, \boldsymbol{V} \in \{0,1\}^{K,R} \quad \text{s.t.} \quad \boldsymbol{V}\boldsymbol{1}_R = \boldsymbol{1}_K \right\}, \tag{4.30}$$

where $\boldsymbol{1}_R$ and $\boldsymbol{1}_K$ are constant vectors only containing ones. We use this parametrization below to minimize the emiprical risk.

Likelihood and Regularization Since the parametrization of the elements of $\boldsymbol{\Theta}_{\text{HC}}$ depends on hidden variables, we introduce a Bayesian formulation with latent classes as in [Hofmann and Puzicha, 1999; Kass and Steffey, 1989] and apply an Expectation-Maximization (**EM**) algorithm to learn the coefficient matrix $\boldsymbol{B}$. To this end, we formulate below the minimization of the error as a maximum-likelihood problem.

Let $\boldsymbol{\mathcal{X}} := (\boldsymbol{X}^{(k)})_{k=1,\dots,K} \in \mathbb{R}^{n,p,K}$ be the *local* design matrices for the K different tasks, with n the number of observations and p the *common* number of features, and let $\boldsymbol{Y} \in \mathbb{R}^{n,K}$ be the target observation matrix. To model the cluster assignment of the substations and write the maximum-likelihood problem, we consider a coefficient matrix $\boldsymbol{U} \in \mathbb{R}^{p,R}$ and a cluster assignment matrix $\boldsymbol{V} \in \{0,1\}^{K,R}$ such that $\boldsymbol{V}\boldsymbol{1}_R = \boldsymbol{1}_K$ like in Equation (4.30). Besides, we introduce an *a priori* discrete probability distribution of the cluster assignments $(\pi_r)_{r=1,\dots,R}$ and the variances within each cluster $(\sigma_r^2)_{r=1,\dots,R} \in \mathbb{R}_+^R$. This is formalized with the following assumptions :

1. The random vectors $(\boldsymbol{v}_k)_{k=1,\dots,K}$ are independent and such that for each $k \in [\![1, K]\!]$, the vector $\boldsymbol{v}_k$ follows a multinomial distribution $\mathcal{M}(1, \pi_1, \dots, \pi_R)$.

2. Conditionally on $v_k^r = 1$, the vector $\boldsymbol{y}^{(k)}$ follows a normal distribution with mean $\boldsymbol{X}^{(k)}\boldsymbol{u}^{(r)}$ and covariance matrix $\sigma_r^2 \boldsymbol{I}_n$. Formally, we have :

$$(\boldsymbol{y}^{(k)} | v_k^r = 1) \quad = \quad \boldsymbol{X}^{(k)}\boldsymbol{u}^{(r)} + \sigma_r \boldsymbol{\varepsilon}^{(k)}, \tag{4.31}$$

where $\boldsymbol{\varepsilon}^{(k)} \in \mathbb{R}^n$ follows the normal distribution $\mathcal{N}(0, \boldsymbol{I}_n)$ with mean 0 and the identity covariance matrix $\boldsymbol{I}_n$.

3. Conditionally on the cluster assignment matrix encoded in the matrix $\boldsymbol{V}$, the random variables $(y_i^k)_{i=1,\dots,n,k=1,\dots,K}$ are independent.

In order to regularize the coefficient vector of each cluster and perform variable selection, we add an Elastic Net penalization [Zou and Hastie, 2005] with hyper-parameters $\lambda > 0$ and $\alpha \in [0,1]$. The regularized maximum-likelihood problem is :

$$\begin{aligned}
&\max_{\boldsymbol{U}\in\mathbb{R}^{p,R},\boldsymbol{\sigma}\in\mathbb{R}_+^R} \mathbb{E}_{\boldsymbol{V}}\left[\log p(\boldsymbol{Y},\boldsymbol{V})\right] - \lambda\left(\alpha\left\|\boldsymbol{U}\right\|_1 + \frac{1-\alpha}{2}\left\|\boldsymbol{U}\right\|_F^2\right) \\
&= \max_{\boldsymbol{U}\in\mathbb{R}^{p,R},\boldsymbol{\sigma}\in\mathbb{R}_+^R} \mathbb{E}_{\boldsymbol{V}}\left[\sum_{k=1}^{K}\sum_{r=1}^{R} v_k^r\left(\log(\pi_r) - \frac{n}{2}\log(2\pi\sigma_r^2) - \frac{1}{2\sigma_r^2}\left\|\boldsymbol{X}^{(k)}\boldsymbol{u}^{(r)} - \boldsymbol{y}^{(k)}\right\|_2^2\right)\right] \\
&\qquad - \lambda\left(\alpha\left\|\boldsymbol{U}\right\|_1 + \frac{1-\alpha}{2}\left\|\boldsymbol{U}\right\|_F^2\right), \qquad (4.32)
\end{aligned}$$

where $p(\boldsymbol{Y},\boldsymbol{V})$ denotes the density function of the pair $(\boldsymbol{Y},\boldsymbol{V})$, the assignments $(v_k^r)_{r=1,\dots,R,k=1,\dots,K}$ are unknown random variables, $\boldsymbol{Y}$ and $\boldsymbol{\mathcal{X}}$ are fixed observations while $(\pi_r)_{r=1,\dots,R}$, $(\boldsymbol{u}^{(r)})_{r=1,\dots,R}$ and $(\sigma_r)_{r=1,\dots,R}$ are unknown parameters.

EM algorithm The E-step of the EM algorithm applied to Problem 4.32 consists in computing for all $r \in [\![1,R]\!]$ and $k \in [\![1,K]\!]$ the *a posteriori* probability of assignment $\gamma_r^k := \mathbb{E}\left[v_k^r|\boldsymbol{Y}\right]$ while the parameters $(\pi_r)_{r=1,\dots,R}$, $\boldsymbol{U}$ and $(\sigma_r^2)_{r=1,\dots,R}$ are kept fixed.

As for the M-step of the EM algorithm, it consists in minimizing with respect to $\boldsymbol{U} \in \mathbb{R}^{p,R}$, $(\sigma_r^2)_{r=1,\dots,R} \in \mathbb{R}_+^R$ and $(\pi_r)_{r=1,\dots,R}$, all other variables being fixed, the regularized expectation :

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{V}}\left[\sum_{k=1}^{K}\sum_{r=1}^{R} \gamma_r^k\left(\log(\pi_r) - \frac{n}{2}\log(2\pi\sigma_r^2) - \frac{1}{2\sigma_r^2}\left\|\boldsymbol{X}^{(k)}\boldsymbol{u}^{(r)} - \boldsymbol{y}^{(k)}\right\|_2^2\right)\right] \\
&\qquad - \lambda\left(\alpha\left\|\boldsymbol{U}\right\|_1 + \frac{1-\alpha}{2}\left\|\boldsymbol{U}\right\|_F^2\right). \qquad (4.33)
\end{aligned}$$

Given the presence of the non-differentiable Elastic Net regularization, a proximal-gradient algorithm is applied. The details of the computations and of the algorithms can be found in [Duchemin, 2018, Section 1.2].

4.4.2 Formulation of the soft clustering problem

Likelihood The exact Equality (4.29) on the coefficient vectors of the substations within a cluster is a relatively strong assumption as it drastically reduces the number of degrees of freedom from pK to $pR + K$. Besides, we see in Section 4.4.3 that empirically, the results with this strict assumption are not fully satisfying. Therefore, we introduce in this section a relaxed version of this constrained model where the columns of the coefficient matrix lie close to one of a few cluster centers :

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T + \boldsymbol{F}, \qquad (4.34)$$

where $\boldsymbol{U}\boldsymbol{V}^T \in \boldsymbol{\Theta}_{\text{HC}}$ and the norm of $\boldsymbol{F}$ is penalized such that $\boldsymbol{B}$ is close to $\boldsymbol{\Theta}_{\text{HC}}$.

Put differently, we assume with the same notations as in Section 4.4.1, that the coefficient vectors $(\boldsymbol{b}^{(k)})_{k=1,\dots,K}$ are gathered around the centers $(\boldsymbol{u}^{(r)})_{r=1,\dots,R}$ of a few clusters such that for each $r \in [\![1, R]\!]$, the within-cluster variance, given by $\sum_{k=1}^{K} v_k^r \left\| \boldsymbol{b}^{(k)} - \sum_{\ell=1}^{K} v_\ell^r \boldsymbol{b}^{(\ell)} \right\|_F^2$, is small.

Formally, in place of Equation (4.31), we assume in this section the following Bayesian model for each task $k \in [\![1, K]\!]$:

$$\left\{ \begin{array}{lcl} (\boldsymbol{b}^{(k)} | v_k^r = 1) & = & \boldsymbol{u}^{(r)} + \tau \boldsymbol{\nu}^{(k)} \\ (\boldsymbol{y}^{(k)} | \boldsymbol{b}^{(k)}) & = & \boldsymbol{X}^{(k)} \boldsymbol{b}^{(k)} + \sigma_k \boldsymbol{\varepsilon}^{(k)}, \end{array} \right. \tag{4.35}$$

where $\tau > 0$ and $(\sigma_k)_{k=1,\dots,K}$ are unknown parameters and for each $k \in [\![1, K]\!]$, the vector $\boldsymbol{\nu}^{(k)} \in \mathbb{R}^p$ follows a normal distribution $\mathcal{N}(0, \boldsymbol{I}_p)$ while the vector $\boldsymbol{\varepsilon}^{(k)} \in \mathbb{R}^n$ follows a normal distribution $\mathcal{N}(0, \boldsymbol{I}_n)$.

With this model, the regularized expected likelihood is :

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{V},\boldsymbol{B}} \left[\log p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{V}) \right] - \lambda \left(\alpha \left\| \boldsymbol{U} \right\|_1 + \frac{1-\alpha}{2} \left\| \boldsymbol{U} \right\|_F^2 \right) \\ = \; & \mathbb{E}_{\boldsymbol{V},\boldsymbol{B}} \left[\sum_{k=1}^{K} \sum_{r=1}^{R} v_k^r \left(\log(\pi_r) - \frac{n}{2} \log(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2} \left\| \boldsymbol{X}^{(k)} \boldsymbol{b}^{(k)} - \boldsymbol{y}^{(k)} \right\|_2^2 \right. \right. \\ & \qquad \left. \left. - \frac{p}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \left\| \boldsymbol{b}^{(k)} - \boldsymbol{u}^{(r)} \right\|_2^2 \right) \right] \\ & - \lambda \left(\alpha \left\| \boldsymbol{U} \right\|_1 + \frac{1-\alpha}{2} \left\| \boldsymbol{U} \right\|_F^2 \right), \end{aligned} \tag{4.36}$$

where $p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{V})$ denotes the density function of the triplet $(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{V})$, the assignments $(v_k^r)_{r=1,\dots,R, k=1,\dots,K}$ and the coefficient matrix $\boldsymbol{B}$ are unknown random variables, $\boldsymbol{Y}$ and $\boldsymbol{\mathcal{X}}$ are fixed observations and $(\pi_r)_{r=1,\dots,R}$, $(\boldsymbol{u}^{(r)})_{r=1,\dots,R}$, τ and $(\sigma_k)_{k=1,\dots,K}$ are unknown parameters. The details of the EM algorithm used to maximize this log-likelihood can be found in [Duchemin, 2018, Section 3.2].

4.4.3 Experiments

GEFCom 2012 The database presented by [Hong et al., 2014] and used in this section contains the measurements of the weather at 11 weather stations and the load of 20 substations in British Columbia with an average load of 82 MWh. In terms of aggregation and load levels, this setting lies between the level of districts and the level of substations described in Table 2.1.

Numerical performances The results for the hard and soft clustering methods are presented in Table 4.1. They are compared with the independent models of Chapter 3. We expected the hard clustering assumption in Equation (4.31) to be too restrictive and indeed, the obtained results are not the best. Although the Bayesian model introduced in Section 4.4.2 leads to an improvement, it does not perform better than the independent models.

The clustering method presented in this section leads to a small but still tangible degradation of the results. The number of parameters was drastically reduced and

	MMr^2	MMMAPE	MRMNMSE
Independent models	0.87	7.11	8.81
Hard clustering in Equation (4.29)	0.852	7.73	9.70
Soft clustering in Equation (4.35)	0.862	7.49	9.34

TABLE 4.1: Results of the clustering methods
Numerical performances on the GEFCom 2012 database of the clustering methods of Section 4.4, compared with the independent models of Chapter 3. While there are 20 load time series in the database, a clustering with 4 groups typically led to the best results with the soft clustering model.

clearly, the clustering assumption that we made is too restrictive and does not lead to a better generalization performance.

Difficulties In terms of optimization, the two main drawbacks of these clustering methods are the slow speed of convergence of the algorithms that we implemented and the existence of local minima. For the latter, we explored three possible solutions, namely a non-random initialization of the algorithm with Ward's method [Ward Jr, 1963], an annealing method [Ueda and Nakano, 1995] to guide the algorithms during the first iterations and resuscitating empty clusters when they occur. In spite of these patches, the algorithms can still converge towards local minima.

The priorities for further investigation seem to be the possibility of clustering only the coefficients corresponding to inputs that are common to all the substations and an improvement of the speed of the algorithms to allow larger experiments.

4.5 Low-rank models

In this section, we also discuss the possibility of imposing a structural constraint on the coefficient matrix in order to reduce the number of parameters of the models. However, instead of the clustering constraints presented in Equation (4.31) and Equation (4.35) that appeared as too restrictive, we introduce in this section a low-rank constraint. This is motivated by the similarity of the tasks enhanced in Section 2.4.2 and Section 4.1, in particular the rapid decrease of the singular values of the coefficient matrices learned with independent models in Figure 4.1.

In order to obtain a low-dimensional representation of the features and improve the generalization performance on different tasks, the low-rank assumption on the coefficient matrix was studied 20 years ago by Intrator and Edelman [1996]. An in-depth theoretical analysis [Ando and Zhang, 2007] later confirmed the interest of this structural assumption and motivated the study of the resulting optimization problems [Bunea et al., 2011, and references therein].

In Section 4.5.1, we introduce a model where the coefficient matrix must be exactly low-rank. In order to select covariates that are relevant to all the tasks, we define in Section 4.5.2 a non-separable regularization that performs joint variable selection. Finally, experiments with the database provided by RTE are discussed in Section 4.5.4.

4.5.1 The low-rank constraint

In the low-rank formulation, we constrain the set of coefficient vectors for the different tasks concatenated into the matrix $\boldsymbol{B} \in \mathbb{R}^{p,K}$, not to be of cardinal $R \in \mathbb{N}^*$ like in Equation (4.29), but to span a subspace of dimension at most R :

$$\operatorname{rank}(\boldsymbol{B}) \leq R, \tag{4.37}$$

where $\operatorname{rank}(\boldsymbol{B})$ denotes the rank of the matrix $\boldsymbol{B}$. Of course, this constraint is effective only if $R < \min(p, K)$.

To highlight the implications of this constraint and how it couples the different tasks, consider $\boldsymbol{B} \in \mathbb{R}^{p,K}$ with $\operatorname{rank}(\boldsymbol{B}) \leq R$ and a pair $(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{p,R} \times \mathbb{R}^{K,R}$ such that :

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T. \tag{4.38}$$

Consider a task $k \in [\![1, K]\!]$ and a vector of covariates $\mathbf{x}^{(k)} \in \mathbb{R}^p$. With the linear model defined in Section 3.2, we have :

$$\hat{\mathsf{y}}_k := \langle \mathbf{x}^{(k)}, \boldsymbol{b}^{(k)} \rangle = \sum_{j=1}^{p} \mathsf{x}_j^k b_j^k = \sum_{r=1}^{R} \langle \mathbf{x}^{(k)}, \boldsymbol{u}^{(r)} \rangle v_k^r, \tag{4.39}$$

where $\boldsymbol{u}^{(r)}$ is the r-th column of $\boldsymbol{U}$ and v_k^r is the element in the k-th row and r-th column of $\boldsymbol{V}$.

The right member of Equation (4.39) is a linear combination of the new covariates $(\langle \mathbf{x}^{(k)}, \boldsymbol{u}^{(1)} \rangle, \ldots, \langle \mathbf{x}^{(k)}, \boldsymbol{u}^{(R)} \rangle)$. In other words, by enforcing the low-rank constraint of Equation (4.37), we force the different tasks to choose collectively a linear transformation of the covariates $\boldsymbol{x} \mapsto (\langle \boldsymbol{x}, \boldsymbol{u}^{(1)} \rangle, \ldots, \langle \boldsymbol{x}, \boldsymbol{u}^{(R)} \rangle)$, characterized by the matrix $\boldsymbol{U} \in \mathbb{R}^{p,R}$, that performs a dimensionality reduction since $R < \min(p, K)$ and such that all tasks can be forecast correctly with the coefficients given by the matrix $\boldsymbol{V} \in \mathbb{R}^{K,R}$.

Therefore, the regularized empirical risk minimization problem that we consider follows from Problem (3.25) with an additional low-rank constraint :

$$\begin{aligned} \min_{\boldsymbol{B} \in \mathbb{R}^{p,K}} \quad & \frac{1}{2n} \sum_{k=1}^{K} \left\| \boldsymbol{y}^{(k)} - \boldsymbol{X}^{(k)} \boldsymbol{b}^{(k)} \right\|_2^2 + \Omega(\boldsymbol{B}), \\ \text{s.t.} \quad & \operatorname{rank}(\boldsymbol{B}) \leq R, \end{aligned} \tag{4.40}$$

with a regularization $\Omega : \mathbb{R}^{p,K} \to \mathbb{R}_+$.

Remark 6. *The low-rank constraint is a strong restriction on the model and just like we introduced the soft version of Equation (4.29) in Equation (4.35), we can define a less restrictive formulation of the low-rank constraint, similarly to [Ando and Zhang, 2007]. Instead of imposing the rank of the coefficient matrix, we may require that it is close to a low-rank matrix i.e. it is the sum of a low-rank matrix with a small perturbation :*

$$\boldsymbol{B} = \boldsymbol{E} + \boldsymbol{F}, \tag{4.41}$$

where $\boldsymbol{E} \in \mathbb{R}^{p,K}$ is such that $\operatorname{rank}(\boldsymbol{E}) \leq R$ and $\boldsymbol{F} \in \mathbb{R}^{p,K}$. To pull the component $\boldsymbol{F}$ towards zero, we may add to the objective a regularization, the regularized empirical

risk minimization problem being in this case :

$$\begin{aligned} \min_{\boldsymbol{E}\in\mathbb{R}^{p,K},\boldsymbol{F}\in\mathbb{R}^{p,K}} \quad & \frac{1}{2n}\sum_{k=1}^{K}\left\|\boldsymbol{y}^{(k)}-\boldsymbol{X}^{(k)}(\boldsymbol{e}^{(k)}+\boldsymbol{f}^{(k)})\right\|_2^2+\Omega(\boldsymbol{E})+\frac{\mu}{2}\left\|\boldsymbol{F}\right\|_F^2, \\ s.t. \quad & rank(\boldsymbol{E})\leq R, \end{aligned} \tag{4.42}$$

with a regularization $\Omega : \mathbb{R}^{p,K} \to \mathbb{R}_+$ *and* $\mu > 0$*. However, it is not clear yet that this leads to interesting empirical results.*

4.5.2 Joint variable selection

In addition to the low-rank constraint set in Problem (4.40), we have considered a group-Lasso regularization [Bakin et al., 1999; Obozinski et al., 2010; Yuan and Lin, 2006] like in Equation (2.2). It is defined for any matrix $\boldsymbol{B} \in \mathbb{R}^{p,K}$ by :

$$\left\|\boldsymbol{B}\right\|_{1,2} := \sum_{j=1}^{p}\left\|\boldsymbol{b}_j\right\|_2 = \sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K}(b_j^k)^2}, \tag{4.43}$$

where $(\boldsymbol{b}_j)_{j\in[\![1,p]\!]}$ are the rows of the matrix $\boldsymbol{B}$. The group-Lasso regularization is known for encouraging some of the groups to have zero norm. Thereby, it induces a common sparsity structure among the tasks. Effectively, that a row $\boldsymbol{b}_j$ of the matrix $\boldsymbol{B}$ with $j \in [\![1, p]\!]$ is zero means that none of the tasks uses the associated covariates $(\mathsf{x}_j^k)_{k=1,\dots,K}$. The minimization problem of the regularized empirical risk is in this case :

$$\begin{aligned} \min_{\boldsymbol{B}\in\mathbb{R}^{p,K}} \quad & \frac{1}{2n}\sum_{k=1}^{K}\left\|\boldsymbol{y}^{(k)}-\boldsymbol{X}^{(k)}\boldsymbol{b}^{(k)}\right\|_2^2+\lambda\left\|\boldsymbol{B}\right\|_{1,2}, \\ \text{s.t.} \quad & \text{rank}(\boldsymbol{B})\leq R, \end{aligned} \tag{4.44}$$

where $\lambda > 0$ is a regularization hyperparameter. Chapter 5 is dedicated to a detailed analysis of this optimization Problem (4.44), with the important additional assumption that the design matrix is the same for all the tasks *i.e.* $\boldsymbol{X}^{(1)} = \ldots = \boldsymbol{X}^{(K)} := \boldsymbol{X} \in \mathbb{R}^{n,p}$:

$$\begin{aligned} \min_{\boldsymbol{B}\in\mathbb{R}^{p,K}} \quad & \frac{1}{2n}\sum_{k=1}^{K}\left\|\boldsymbol{y}^{(k)}-\boldsymbol{X}\boldsymbol{b}^{(k)}\right\|_2^2+\lambda\left\|\boldsymbol{B}\right\|_{1,2}, \\ \text{s.t.} \quad & \text{rank}(\boldsymbol{B})\leq R. \end{aligned} \tag{4.45}$$

4.5.3 Partially low-rank models

In Section 4.5.1, we have considered a low-rank constraint on the whole matrix $\boldsymbol{B} \in \mathbb{R}^{p,K}$. In effect, it constrains the coefficients associated to all the features. Yet, it appears also legitimate to constrain only the coefficients corresponding to the inputs that are shared by all the tasks (*e.g.* the hour of the week and not the past loads). In the experiments of Section 4.5.4, the constraint that we use is even more detailed as we consider disjoint blocks of rows $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_\ell$ of the matrix $\boldsymbol{A}$ and

we impose independent rank constraints on the different blocks. This is formulated as the following optimization problem with partial low-rank constraints :

$$\begin{aligned} \min_{\boldsymbol{A}\in\mathbb{R}^{p_0,K},\boldsymbol{C}\in\mathbb{R}^{p-p_0,K}} \quad & \frac{1}{2n}\sum_{k=1}^{K}\left\|\boldsymbol{y}^{(k)}-\boldsymbol{X}^{(0)}\boldsymbol{a}^{(k)}-\boldsymbol{Z}^{(k)}\boldsymbol{c}^{(k)}\right\|_2^2+\Omega(\boldsymbol{A},\boldsymbol{C}), \\ \text{s.t.} \quad & \operatorname{rank}(\boldsymbol{A}_1)\leq R_1,\ldots,\operatorname{rank}(\boldsymbol{A}_\ell)\leq R_\ell, \end{aligned} \tag{4.46}$$

where $R_1,\ldots,R_\ell\in\mathbb{N}$.

4.5.4 Experiments with partially low-rank models

So far, we have not obtained satisfying empirical results with a rank constraint on the whole matrix like in Equation (4.40). Although the difference between the results with and without a low-rank constraint on the entire matrix of coefficients is less significant when working with middle-term models, we have concluded that this constraint is not relevant.

Instead, we focus directly on the partially low-rank models of Equation (4.46). Again, we could not improve the generalization performance of the local models with low-rank constraints. Still, we have considerably reduced the number of degrees of freedom for a minor degradation of the performance.

Indeed, in Figure 4.5, we compare the performances of the local models without any rank constraints with the partially low-rank problem where the block of coefficients related to the hour of the week is constrained to be of rank r_{h} and the block related to the day of the year is constrained to be of rank r_{d}. The best results with the constraints is obtained for $r_{\mathsf{h}}=r_{\mathsf{d}}=20$. Given that for each of the $\mathcal{K}=1751$ substations there are $p_{\mathsf{h}}=168$ coefficients for the hour of the week and $p_{\mathsf{d}}=32$ for the day of the year, the unconstrained model has about $(p_{\mathsf{h}}+p_{\mathsf{d}})\mathcal{K}=300\,000$ degrees of freedom for these inputs and the constrained model has approximatively $(p_{\mathsf{h}}+\mathcal{K})r_{\mathsf{h}}+(p_{\mathsf{d}}+\mathcal{K})r_{\mathsf{d}}=40\,000$.

The penalizations that we have used to obtain Figure 4.5 are the same as in Chapter 3. Unfortunately, the group-Lasso regularization that we have introduced in Equation (4.44) does not lead to better results. It was effectively introduce to consider a potential variable selection procedure, which is not the case here since first, the past loads and the weather stations have already been selected in Section 3.6.5 and secondly, the family of features introduced for each input in Section 3.1 is not redundant.

While it is disappointing not to improve the generalization performance, this result still proves that the number of degrees of freedom in the independent models is unnecessary large. Besides, the analysis of the estimated low-rank matrix is a potential way of understanding the underlying structure.

4.6 Sum consistent local models

In Section 4.4 and Section 4.5, we have imposed specific structures on the coefficient matrix with constraints and regularizations. These structures belong to the *task structure* category proposed by Rai et al. [2012]. In this section, we focus on the *output structure* and couple the tasks by introducing a multi-objective optimization

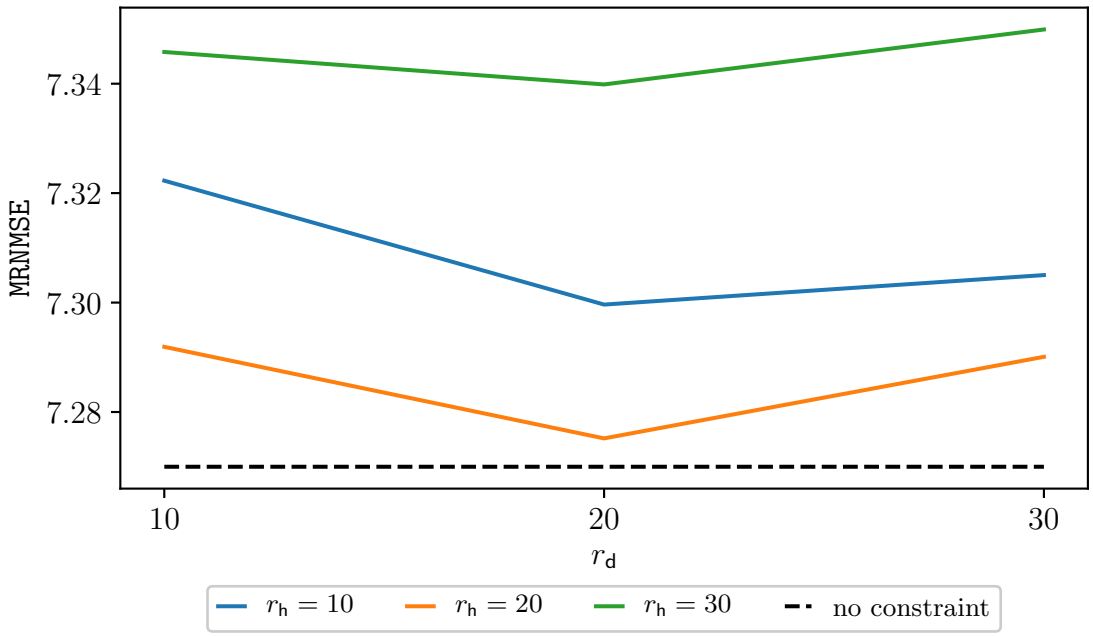


FIGURE 4.5: Results with the low-rank model
MRMNMSE on the test year 2016 for different values of the rank r_h of the block of coefficients related to the hour of the week and different values of the rank r_d of the block of coefficients related to the day of the year. The penalizations are the same as in Chapter 3 and the hyperparameters have been optimized for each configuration.

problem defined by considering two levels : one for the individual substations and one for aggregated loads.

There are different possibilities to leverage a multi-level structure in a forecasting problem. First, the Hierarchical Linear Models (**HLM**) provide a framework designed especially for problems with nested groups. Secondly, with a two-step forecasting and aggregating procedure based on specialized experts, Hyndman et al. [2011] proposed an alternative model consistent both at the local and the aggregated levels. Thirdly, a multivariate problem similar to the sum consistent model of this section has been studied by the mathematical finance community. Avellaneda and Boyer-Olson [2002]; Cont and Deguest [2013]; Durrleman and El Karoui [2008]; Jourdain and Sbai [2012] studied different approaches to ensure the consistency between observed index options (value of a basket of options) and options on index components (values of the individual options). The model that we consider here for load forecasting is inspired from this third possibility, the analogous of a basket of options being the aggregated load of a group of substations.

4.6.1 Multi-objective loss

The model that we consider is similar to the models defined in Chapter 3, only the loss is different. The proposed sum consistent loss has two components : it still includes the part that measures the error for each load time series, it also measures

the errors made on the aggregated time series. Effectively, we add an augmented data-fitting term to the squared Frobenius norm that measures the individual errors. Let $\boldsymbol{M} \in \mathbb{R}^{K,K}$, $\mu > 0$ and $(\mathbf{y}, \hat{\mathbf{y}}) \in \mathbb{R}^{K,2}$, we define the sum-consistent loss as :

$$\mathcal{L}_{\boldsymbol{M},\mu}(\mathbf{y}, \hat{\mathbf{y}}) := \frac{1}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 + \frac{\mu}{2} \|\boldsymbol{M}(\mathbf{y} - \hat{\mathbf{y}})\|_2^2 . \tag{4.47}$$

The matrix $\boldsymbol{M}$ is simply a fixed parameter and not a representation of the covariance of the outputs, its goal is to ensure that the local forecasts are consistent with the loads aggregated at a higher level.

In Example 7 and Example 8, we show how, choosing $\boldsymbol{M}$ appropriately, the sum consistent loss can measure the error committed on aggregated normalized loads. Thereby, the minimization of $\mathcal{L}_{\boldsymbol{M},\mu}$ on a datasets leads to a compromise between the local errors and the aggregated errors.

Example 7. *Consider for example* $\boldsymbol{M} = \frac{\mathbf{1}_K \mathbf{1}_K^T}{K}$, *in this case we have :*

$$\frac{1}{2} \|\boldsymbol{M}(\mathbf{y} - \hat{\mathbf{y}})\|_2^2 = \frac{K}{2} \left\| \frac{1}{K} \sum_{k=1}^{K} (\mathsf{y}_k - \hat{\mathsf{y}}_k) \right\|_F^2 . \tag{4.48}$$

The augmented data-fitting term in Equation (4.48) encourages the model to make an accurate prediction of the normalized and uniformly aggregated load $\sum_{k=1}^{K} \mathsf{y}_k$.

Example 8. *A second example of interest for the substations level, that is the most interesting empirically, is the case where* $(Z_g)_{g \in [\![1,G]\!]}$ *is a partition of the* K *substations into* $G \in \mathbb{N}^*$ *groups and for all* $i, j \in [\![1, K]\!]$, *we set :*

$$\boldsymbol{M}_{i,j} = \sum_{g=1}^{G} \frac{1_{i \in Z_g} 1_{j \in Z_g}}{\eta_g}, \tag{4.49}$$

where $\eta_g := |Z_g|$ *is the number of substations in the group* Z_g. *Up to a permutation of the tasks, the matrix* $\boldsymbol{M}$ *is block diagonal. Let* $\boldsymbol{J}_g := \frac{1}{\eta_g} \mathbf{1}_{\eta_g} \mathbf{1}_{\eta_g}^T$, *we have :*

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{J}_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \boldsymbol{J}_G \end{bmatrix}.$$

In this case, the augmented data-fitting term is :

$$\sum_{g \in [\![1,G]\!]} \frac{\eta_g}{2} \left\| \frac{1}{\eta_g} \sum_{k \in Z_g} (\mathsf{y}_k - \hat{\mathsf{y}}_k) \right\|_2^2 . \tag{4.50}$$

It is a weighted sum of the squared errors committed on the aggregated normalized loads $\sum_{k \in Z_g} \mathsf{y}_k$, *in each group* Z_g *with* $g \in [\![1, G]\!]$, *by the aggregated forecasts* $\sum_{k \in Z_g} \hat{\mathsf{y}}_k$. *Note that in this case, the problem is separable over the different groups.*

4.6.2 Motivation for the multi-level consistency

We have illustrated in Figure 4.2 and Figure 4.3 the important correlations between the residuals at the local level. If the residuals tend to all point in the same directions, then they should accumulate and lead to tangible errors at the aggregated levels, which would be penalized by the sum consistent loss.

The purpose of the augmented data-fitting term is precisely to provide the model with an additional (soft) constraint to help it being consistent, that is to say, making forecasts that simultaneously have low local errors and reasonable errors on the aggregated loads, altogether for a better generalization. Besides, it provides the TSO with a guarantee that forecasts in the corresponding parts of the network represent a reasonable prediction of the aggregated demands, at the national level if we choose Example 7 and at the levels of the groups with Example 8.

To assess its potential, we compare in Table 4.2 the errors committed on the aggregated loads by two groups of forecasting models, all defined like in Chapter 3 with the parametrization given in Tables 3.2, 3.3 and 3.4. The performances given in the column *Independent forecasts* of Table 4.2 are measured on the aggregated loads and the models for this column were precisely trained by minimizing the errors on these aggregated loads. For the right column, there is only one model, that is trained by minimizing the errors at the level of the substations. The forecasts that it produces are then aggregated so that the performances given in the column *Aggregated local forecasts* of Table 4.2 also correspond to the aggregated loads, while they were estimated independently and only with the local loads.

Because the performances on the right column are not as good as the performances in the middle column, we believe that there is room for improvement and the sum consistent loss can potentially guide the local models and help them to make more consistent forecasts.

Aggregation level	Independent forecasts	Aggregated local forecasts
National	(0.983, 1.31, 1.69)	(0.974, 1.55, 2.27)
RTE regions	(0.959, 2.07, 2.71)	(0.962, 1.99, 2.83)
Administrative	(0.963, 2.07, 2.66)	(0.957, 2.15, 3.06)
Districts	(0.954, 2.27, 2.93)	(0.947, 2.38, 3.33)

TABLE 4.2: Aggregated-load forecasts and aggregated forecasts ($\texttt{MMr}^2$, $\texttt{MMMAPE}$, $\texttt{MRMNMSE}$) as defined in Section 2.7 to compare the independent short-term models defined in Chapter 3 with the parametrization given in Tables 3.2, 3.3 and 3.4, estimated and evaluated with loads at the same levels, with the forecasts obtained by aggregating the outputs of the local models. The models presented in this table are all estimated by minimizing a separable squared error term.

4.6.3 Results with the sum consistent loss

We focus on the load forecasting problem at the level of substations with the sum consistent loss of Equation (4.47). We have considered different possibilities for the experiments : namely a matrix $\boldsymbol{M}$ such that the error on the nationally aggregated

load is penalized, a matrix $\boldsymbol{M}$ such that for each substation, the errors on the load aggregated with the $\nu \in \mathbb{N}^*$ nearest neighbor is penalized, and finally a matrix $\boldsymbol{M}$ defined like in Example 8, where the zones correspond to the districts defined in Section 2.5.2. With the two first settings, we observed an undeniable degradation of the forecasts at the level of substations. The last setting on the other hand, is the most satisfactory and we illustrate the results below.

In other words, we minimize the sum of two terms : the errors committed at the level of the substations and the errors on the load aggregated in each districts. Let $(Z_g)_{g\in[\![1,G]\!]}$ denote the partition into the $G = 32$ districts of the $K = 1751$ substations. The corresponding loss is given by :

$$\mathcal{L}_{\boldsymbol{M},\mu}(\mathbf{y},\hat{\mathbf{y}}) := \frac{1}{2}\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|_2^2 + \frac{\mu}{2}\sum_{g\in[\![1,G]\!]}\eta_g\left\|\frac{1}{\eta_g}\sum_{k\in Z_g}(\mathsf{y}_k-\hat{\mathsf{y}}_k)\right\|_2^2, \tag{4.51}$$

where $\mathbf{y} \in \mathbb{R}^K$ contains the loads of the substations, $\hat{\mathbf{y}} \in \mathbb{R}^K$ is its estimate and $\mu \geq 0$.

In short, we observed that :

- The forecasts of the aggregated loads are improved in all the districts.
- The `RMNMSE` defined for the whole set of substations is not better.
- The `RMNMSE` restricted to individual districts is not always improved, as illustrated in Figure 4.6.
- In some districts, the individual forecasts benefit from the sum consistent loss and both the aggregated and the individual forecasts are improved, as illustrated in Figure 4.7.

Out of the 32 districts, the sum consistent loss lead to better individual forecasts in 10 districts, like in Figure 4.7. In the 22 others, we have observed a degradation of the numerical performances with the sum consistent loss.

These results let us believe that the districts are a relevant scale to couple the different tasks while a coupling at the national level does not lead to better performances. With the negative results in some districts however, we tend to conclude that a coupling is not relevant for all the substations and a procedure to select which substations should be coupled would be helpful.

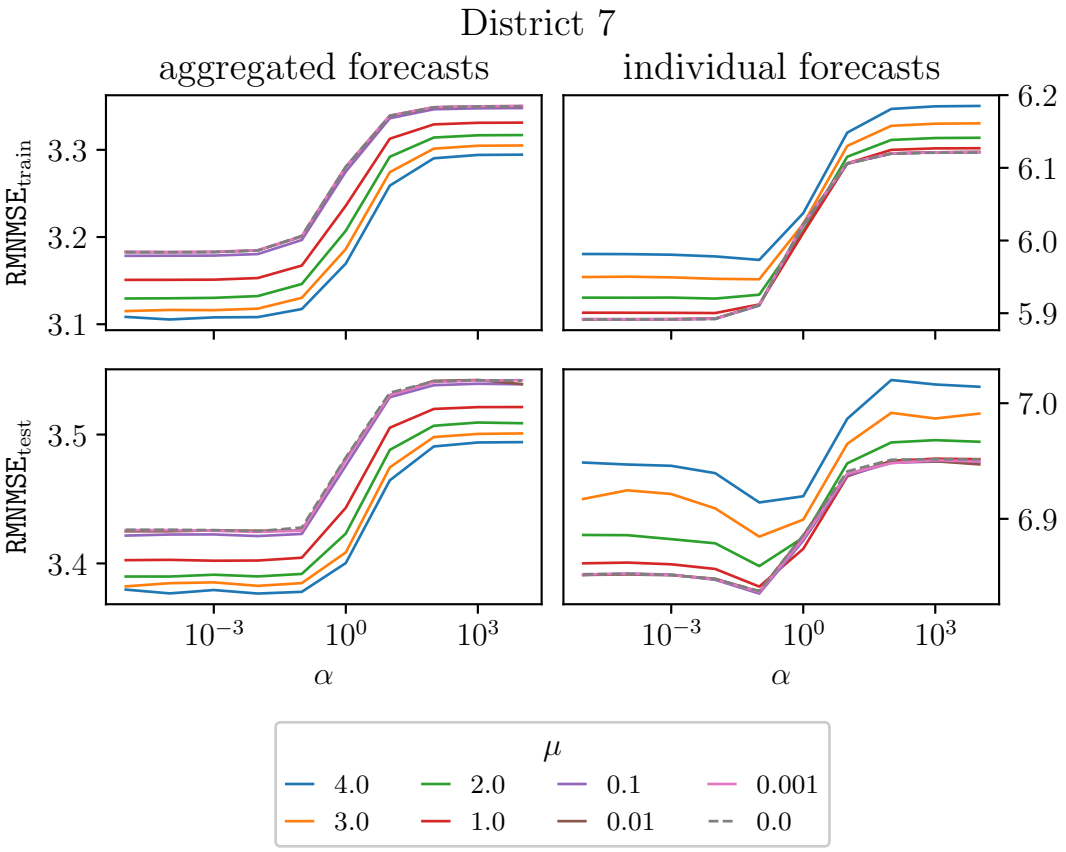


FIGURE 4.6: Failure with the sum consistent loss in District 7
The coefficient μ determines the compromise in Equation (4.51) between the attention paid to the individual forecasts and the aggregated forecasts. The coefficient α is used to multiply all the regularization hyperparameters of the models : $\alpha < 1$ corresponds to a weaker regularization and $\alpha > 1$ corresponds to a stronger regularization.

(*top left*) `RNMSE` on the training set of the forecasts aggregated in District 7. It measures the error $\left\| \frac{1}{\eta_7} \sum_{k \in Z_7} (y_k - \hat{y}_k) \right\|_2^2$

(*bottom left*) `RNMSE` on the test set of the forecasts aggregated in District 7.

(*top right*) `RMNMSE` on the training set of the individual forecasts in District 7. It measures the error $\frac{1}{\eta_7} \sum_{k \in Z_7} (y_k - \hat{y}_k)^2$

(*bottom left*) `RMNMSE` on the test set of the individual forecasts in District 7.

With the training and the test sets, the variations of the error measure are monotone when μ increases, it consistently decreases at the aggregated level as expected but increases for the individual forecasts. The sum consistent loss does not help the model to make better local predictions.

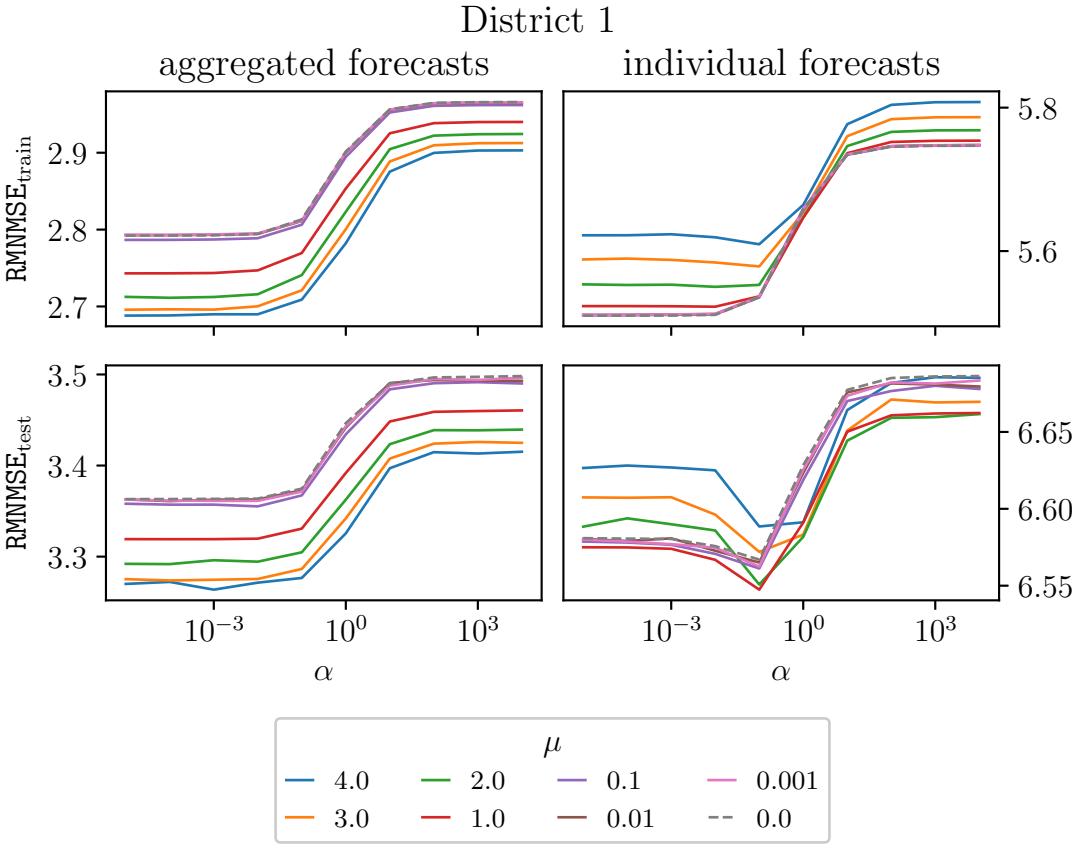


FIGURE 4.7: Success with the sum consistent loss in District 1

(*top left*) RNMSE on the training set of the forecasts aggregated in District 1. It measures the error $\left\| \frac{1}{\eta_1} \sum_{k \in Z_1} (\mathsf{y}_k - \hat{\mathsf{y}}_k) \right\|_2^2$.

(*bottom left*) RNMSE on the test set of the forecasts aggregated in District 1.

(*top right*) RMNMSE on the training set of the individual forecasts in District 1. It measures the error $\frac{1}{\eta_7} \sum_{k \in Z_1} (\mathsf{y}_k - \hat{\mathsf{y}}_k)^2$.

(*bottom left*) RMNMSE on the test set of the individual forecasts in District 1.

With the training set, the variations of the RMNMSE when μ increases are monotone, it consistently decreases at the aggregated level and increases for the individual forecasts, as expected. With the test set, the error on the aggregated load decreases with μ as well meaning that the aggregated load is better predicted : it is a guarantee for the TSO that in the district, the local forecasts are more consistent with the demand of the whole district. The most interesting graph is at the bottom right. The addition of the augmented data-fitting term in the sum consistent loss helped the model to make better forecasts at the level of the substations.

4.7 Conclusion of Chapter 4

From the standard bivariate modeling described in Chapter 3, we have considered three different possibilities in Chapter 4 to couple the models in a multi-task framework.

The clustering assumption made in Section 4.4 and the low-rank assumption in Section 4.5 are quite similar, they both consider that the coefficient matrix with $p \times K$ coefficients can be parametrized with less degrees of freedom. The former has not lead to an improvement and we conclude that, as presented, this assumption is too restrictive. However, the latter let us conclude that the number of parameters in the independent models described in Chapter 3 is indeed unnecessarily large.

In Section 4.6, we have adopted a different approach where the number of parameters in the model is not changed but the forecasts for the individual substations must be consistent with the observations at the districts level. The positive results in Figure 4.7 show that an improvement of the local forecasts is possible with such a coupling of the local models. In particular, we believe that coupling the models for the substations in the same district is more relevant than coupling all the substations in France. This is consistent with the intuition that we can have with Figure 4.2. Still, the negative results in Figure 4.6 show the importance of selecting which substations can benefit from a coupling and identify which substations have an outlying behavior and do not benefit from a multi-task approach.

Chapter 5

Fast algorithms for Sparse Reduced Rank Regression

This chapter is dedicated to the study of the sparse and low-rank regression Problem (4.45). The content is extracted from the article [Dubois et al., 2019] accepted at the International Conference on Artificial Intelligence and Statistics in 2019. The notations are specific to this chapter.

Abstract We consider a reformulation of Reduced-Rank Regression (RRR) and Sparse Reduced-Rank Regression (SRRR) as a non-convex non-differentiable function of a single of the two matrices usually introduced to parametrize low-rank matrix learning problems. We study the behavior of proximal gradient algorithms for the minimization of the objective. In particular, based on an analysis of the geometry of the problem, we establish that a *proximal* Polyak-Łojasiewicz inequality is satisfied in a neighborhood of the set of optima under a condition on the regularization parameter. We consequently derive linear convergence rates for the proximal gradient descent with line search and for related algorithms in a neighborhood of the optima. Our experiments show that our formulation leads to much faster learning algorithms for RRR and especially for SRRR.

5.1 Introduction

In matrix learning problems, an effective way of reducing the number of degrees of freedom is to constrain the rank of the coefficient matrix to be learned. Low-rank constraints lead however to non-convex optimization problems for which the structure of critical points and the behavior of standard optimization algorithms, like gradient descent, stochastic block coordinate gradient descent and their proximal counterparts, are difficult to analyze. These difficulties have lead researchers to either use these algorithms without guarantee or to consider convex relaxations in which the low-rank constraint is replaced by a trace-norm constraint or penalty. In

the last few years however, a better understanding of the geometry of these problems [Li et al., 2016; Zhu et al., 2017b], new tools from non-convex analysis [Attouch and Bolte, 2009; Csiba and Richtarik, 2017; Frankel et al., 2015; Karimi et al., 2016; Khamaru and Wainwright, 2018] as well as results on the behavior of standard algorithms around saddle points [Lee et al., 2017] were developed under regularity assumptions to analyze their convergence and eventually prove rates of convergence.

Formulations that require to learn a low-rank matrix or its factors appear in many problems in machine learning, from variants of Principal Components Analysis and Canonical Correlation Analysis, to matrix completion problems and multi-task learning formulations. Reduced-Rank Regression (**RRR**) is a fundamental model of this family. It corresponds to the multiple outputs linear regression in which all the vectors of parameters associated with the different dimensions are constrained to lie in a space of dimension $r \in \mathbb{N}^*$. Precisely, if $X \in \mathbb{R}^{n,p}$ is a design matrix and $Y \in \mathbb{R}^{n,k}$ has columns corresponding to the multiple tasks, then the problem is usually formulated with $\|\cdot\|_F$ the Frobenius norm as

$$\min_{W \in \mathbb{R}^{p,k}:\ \mathrm{rank}(W) \leq r} \quad \frac{1}{2} \|Y - XW\|_F^2 \,. \tag{5.1}$$

The solution of Problem (5.1) can be obtained in closed form [Velu and Reinsel, 2013] and requires to project the usual multivariate linear regression parameter estimate on the subspace spanned by the top right singular vectors of the matrix $(X^T X)^{-1/2} X^T Y$.

Sparse Reduced-Rank Regression (**SRRR**) is a variant in which the objective is regularized by the group-Lasso norm $\|W\|_{1,2} = \sum_i (\sum_j W_{ij}^2)^{1/2}$, in order to induce row-wise sparsity in the matrix W, which corresponds to simultaneous variable selection for all tasks. Given $\lambda > 0$, the optimization problem takes the form :

$$\min_{W \in \mathbb{R}^{p,k}:\ \mathrm{rank}(W) \leq r} \frac{1}{2} \|Y - XW\|_F^2 + \lambda \|W\|_{1,2} \,. \tag{5.2}$$

For this formulation, there is no closed form solution anymore, and the conceptually simple algorithms that have been proposed to solve Problem (5.2) are not so computationally efficient.

In the last decade, many optimization problems of the form :

$$\min_{W \in \mathbb{R}^{p,k}:\ \mathrm{rank}(W) \leq r} \quad \mathcal{F}(W), \tag{5.3}$$

with $\mathcal{F}$ a convex function have been tackled via the convex relaxation obtained by replacing the rank constraint with a constraint or a regularization on the trace-norm $\|W\|_*$. unfortunately, these formulations often lead to expensive algorithms and the relaxation induces a bias. A recent literature revisited a number of these problems based on an explicit parameterization of the low-rank matrix, as biconvex problems of the form :

$$\min_{U \in \mathbb{R}^{p,r},\, V \in \mathbb{R}^{k,r}} \quad \mathcal{F}(UV^T). \tag{5.4}$$

In particular, it is natural to formulate Problem (5.1) and Problem (5.2) in this form.

In this paper, we additionally impose $V^TV = I_r$ without loss of generality and we reformulate the SRRR problem as a non-convex non-differentiable optimization problem of a single thin matrix U. Based on the geometry of the objective described in Corollary 14, we establish in Corollary 17 a generalized Polyak-Łojasewicz inequality [Karimi et al., 2016; Polyak, 1963] in a neighborhood of the minima which can be leveraged to show in Corollary 18 asymptotic linear convergence of the proximal gradient algorithm and of stochastic block coordinate proximal descent algorithms. Our results are also relevant to solve very large-scale RRR instances for which the direct computation of the closed form solution would not be possible.

The paper is structured as follows. In Section 5.2, we discuss related work. In Section 5.3, we reformulate the RRR/SRRR problems. In Section 5.4, we obtain global convergence results. To analyze the local convergence in Section 5.5, we review the structure of RRR and establish properties based on the orthogonal invariance of the objective as well as the convexity of its restriction on certain cones in a neighborhood of the optima. Thus, we obtain a Polyak-Łojasiewicz inequality and a generalized Polyak-Łojasiewicz inequality respectively for RRR and SRRR in a neighborhood of the global minima. Finally, Section 5.6 illustrates with numerical experiments the performances of the proposed algorithms.

5.2 Related Work

Velu and Reinsel [2013] studied Problem (5.1) and showed that it is one of the few low-rank matrix problems which has a closed form solution. Baldi and Hornik [1989] studied thoroughly the biconvex version of Problem (5.1) and identified its critical points to show that its local minima are global. Bunea et al. [2011, 2012]; Chen and Huang [2012]; Ma and Sun [2014]; Mukherjee et al. [2015]; She [2017] considered Problem (5.2) and highlighted the statistical properties of the estimator. The algorithms proposed in these papers all consist essentially in optimizing alternatingly with respect to U and V an objective of the form (5.4) (and more precisely the objective (5.5) introduced in Section 5.3) under the constraint $V^TV = I_r$. The full optimization w.r.t. V requires to compute an SVD of the matrix $Y^TXU \in \mathbb{R}^{k,r}$ which is of reasonable size, but the full optimization w.r.t. U requires to solve a full group-Lasso problem.

Among others, iterative first-order algorithms that are classical for the jointly convex setting may be applied to the non-convex Problem (5.4). Until recently, precise convergence guarantees were relatively rare but the observation of good empirical rates of convergence motivated a finer analysis. In particular, a number of recent papers established stronger theoretical results for these algorithms in the smooth non-convex case. Notably, Jain et al. [2017] obtained the first *global* linear rate of convergence for the very particular case of the matrix square-root computation. For more general biconvex formulations, Park et al. [2016] and Wang et al. [2016] established convergence rate guarantees for the gradient descent algorithm for Problem (5.4) provided an appropriate initialization is used and penalties such as $\frac{1}{4}\|U^TU - V^TV\|_F^2$ are added to the objective as regularizers.

As a consequence of the aforementioned performances, there was a regain of interest for the biconvex problems like (5.4) and their geometry has been studied in numerous papers. Bhojanapalli et al. [2016]; Boumal et al. [2016]; Ge et al. [2017,

2016]; Kawaguchi [2016]; Li et al. [2017, 2018]; Zhu et al. [2017a] studied critical points and made use of the strict saddle property to show global convergence results for gradient descent and stochastic variants. Some of these works define a partition of the space and characterize the behavior of gradient descent in each region [Li et al., 2016; Zhu et al., 2017b].

Besides, it was shown recently that appropriate first-order algorithms cannot converge to saddle points when the curvature of the objective is strict around them [Lee et al., 2017; Panageas and Piliouras, 2016; Sun et al., 2015]. These algorithms actually spend only a limited amount of time near the saddle points if the Hessian is Lipschitz [Du et al., 2017; Jin et al., 2017]. However, these papers do not provide general convergence rate results, in particular not in the non-differentiable case.

From the performances of classical first-order algorithms originated attempts to characterize convergence and to possibly prove rates based on the local geometry of non-convex objective functions around minima. In particular, Karimi et al. [2016] reviewed and provided a unified point of view of the recent literature on the Polyak-Łojasiewicz inequality [Polyak, 1963]. This type of results was leveraged by Csiba and Richtarik [2017] to prove convergence rates. A parallel thread of research focused on the Kurdyka-Łojasiewicz inequality (**KŁ**), with the motivation that all semi-algebraic functions satisfy it. Attouch and Bolte [2009]; Attouch et al. [2013]; Frankel et al. [2015]; Ochs et al. [2014] were able to characterize asymptotic convergence rates for the forward-backward algorithm under the KŁ inequality. These types of results were extended for block coordinate descent schemes in Attouch et al. [2010]; Bolte et al. [2014]; Nikolova and Tan [2017]; Xu and Yin [2017], and for accelerated proximal descent algorithms in Chouzenoux et al. [2014]; Li and Lin [2015]. However, in general, it remains difficult to prove a specific rate for a given problem, because the exact rate depends on the best exponent that can be obtained in the KŁ inequality, and with the exception of some results provided in Li and Pong [2017], determining this exponent remains difficult.

5.3 Reformulation and algorithm

5.3.1 New formulation for RRR/SRRR with one thin matrix U

We reformulate the biconvex version of SRRR :

$$\min_{U\in\mathbb{R}^{p,r},\, V\in\mathbb{R}^{k,r}} \quad \frac{1}{2}\left\|Y - XUV^T\right\|_F^2 + \lambda \left\|UV^T\right\|_{1,2}, \tag{5.5}$$

by eliminating V as follows. First, we can impose $V^TV = I_r$ as in Chen and Huang [2012] without loss of generality. Then, expanding the Frobenius norm and using the invariance of the norms to the transformation $U \mapsto UV^T$ with $V \in \mathbb{R}^{k,r}$ such that $V^TV = I_r$, the objective becomes $\frac{1}{2}\|XU\|_F^2 - \langle Y, XUV^T\rangle + \lambda \|U\|_{1,2}$ where $\langle\cdot,\cdot\rangle$ is the Frobenius inner product. The value of the orthogonal Procrustes problem :

$$\max_{V\in\mathbb{R}^{k,r}:\, V^TV=I_r} \quad \langle Y, XUV^T\rangle,$$

is the trace-norm $\left\|Y^TXU\right\|_*$ (*cf.* Fact 33 in Appendix G.3). So, letting $f(U) := f_1(U) - f_2(U)$ with :

$$f_1(U) = \frac{1}{2}\left\|XU\right\|_F^2 \quad \text{and} \quad f_2(U) = \left\|Y^TXU\right\|_*,$$

and $F^\lambda(U) := f(U) + \lambda \left\|U\right\|_{1,2}$, RRR and SRRR are respectively reformulated as :

$$\min_{U\in\mathbb{R}^{p,r}} f(U), \tag{RRR}$$

$$\min_{U\in\mathbb{R}^{p,r}} F^\lambda(U). \tag{SRRR}$$

The objectives, as differences of convex functions, are clearly non-convex. However, they are still orthogonal-invariant *i.e.* for any $U \in \mathbb{R}^{p,r}$ and $R \in \mathbb{R}^{r,r}$ such that $R^TR = I_r$, we have $f(UR) = f(U)$ and $F^\lambda(UR) = F^\lambda(U)$. Note that the above derivations would still be valid if we replaced the row-wise group-Lasso $\left\|\cdot\right\|_{1,2}$ by any regularizer which is invariant when the argument is multiplied on the right by an orthogonal matrix.

Also, note that although f involves a trace-norm, its argument, Y^TXU, is of dimensions $k \times r$ while, in convex relaxations of low-rank formulations like Problem (5.3), the rank constraint is substituted with a trace-norm regularizer $\left\|W\right\|_*$ that is computed for a typically large matrix W of dimensions $p \times k$.

5.3.2 Optima of the classical RRR formulation

Velu and Reinsel [2013] characterized the closed form solution of Problem (5.1) when X^TX is invertible as follows. Let $W^* := (X^TX)^{-1}X^TY$ denote the full-rank least squares estimator. Let PSQ^T be the reduced singular value decomposition of $(X^TX)^{-\frac{1}{2}}X^TY$. If the latter has rank ℓ then $P \in \mathbb{R}^{p,\ell}$ and $Q \in \mathbb{R}^{k,\ell}$ have orthonormal columns and $S \in \mathbb{R}^{\ell,\ell}$ is the diagonal matrix with singular values $s_1 \geq \ldots \geq s_\ell > 0$. The solution of Problem (5.1) is unique if $s_r > s_{r+1}$: let $Q_r \in \mathbb{R}^{k,r}$ be the matrix obtained by keeping the first r columns of Q, the solution is $W_r^* := W^*Q_rQ_r^T$.

5.3.3 Algorithms and complexity

The algorithms that we consider are essentially proximal gradient algorithms with line search, except for the fact that f_2 is not differentiable when Y^TXU is not full-rank, which entails that f is not differentiable everywhere. To address this issue, and given that f is a difference of a smooth convex function and a continuous convex function, we consider the subgradient-type algorithms proposed in Khamaru and Wainwright [2018].

Given $U \in \mathbb{R}^{p,r}$, the idea is to use a subgradient z_U of f_2. We assume that X^TX is invertible but consider a more general case in Appendix G.4.1.2 where we detail the computations. Given $R_1DR_2^T$ a singular value decomposition of Y^TXU such that Im $R_1 \subset$ Im Y^TX, we compute $z_U = X^TYR_1R_2^T$ with $R_1 \in \mathbb{R}^{k,r}$, $R_1^TR_1 = I_r$, $D = \text{diag}(d_1 \geq \ldots \geq d_r) \in \mathbb{R}^{r,r}$ with $d_r \geq 0$ and $R_2 \in \mathcal{O}_r$. With a slight abuse of notation, we define $\nabla f(U) := \nabla f_1(U) - z_U$. Note that this is the gradient of

the natural DC programming upper bound. We introduce for any $t > 0$ the t-approximation functions of f and F^λ at U :

$$\tilde{f}_{t,U}(U') := f(U) + \langle \nabla f(U), U' - U \rangle + \frac{1}{2t} \|U' - U\|_F^2 ,$$

and :

$$\tilde{F}^\lambda_{t,U}(U') := \tilde{f}_{t,U}(U') + \lambda \|U'\|_{1,2} .$$

At each iteration of Algorithm 1, the matrix U is updated with Algorithm 2 to U_+ the unique minimizer of $\tilde{F}^\lambda_{t,U}$ if the line search condition :

$$\tilde{F}^\lambda_{t,U}(U_+) \geq F^\lambda(U_+), \tag{LS}$$

is satisfied. Otherwise, t is decreased by a multiplicative factor $\beta < 1$. We explain why Algorithm 2 terminates in Appendix G.5.2. The obtained algorithm is almost a gradient descent algorithm when $\lambda = 0$ and a proximal gradient descent algorithm when $\lambda > 0$ (see Appendix G.4.2). In practice, our algorithms stay away from points where f is non-differentiable and reduce to plain gradient descent and plain proximal gradient descent respectively. This motivated us to also consider for the experiments the accelerated proximal gradient algorithm of Li and Lin [2015], designed for the non-convex setting. We adapt in Section 5.4 parts of the global convergence results of Khamaru and Wainwright [2018] to our algorithms.

Algorithm 1 Proximal Gradient Descent with LSP

Input: data X, Y, $\bar{t}$, starting point $\bar{U}$
Initialize $k = 0$, $U_0 \leftarrow \bar{U}$, $t_{-1} \leftarrow \bar{t}$
while not converged **do**
  Compute t, U_+ with t_{k-1}, U_k and Algorithm 2
  $t_k \leftarrow t$
  $U_{k+1} \leftarrow U_+$
  $k = k + 1$
end while

Algorithm 2 Line Search Procedure (LSP)

Input: t_{k-1}, U_k, parameters $\beta \in (0, 1)$, $\pi \in (0, 1]$
Set $t \leftarrow \frac{t_{k-1}}{\beta}$ with probability π, otherwise $t \leftarrow t_{k-1}$
$U_+ \leftarrow \operatorname{argmin}_{U'} \tilde{F}^\lambda_{t,U_k}(U')$
while (LS) is not satisfied **do**
  $t \leftarrow \beta t$
  $U_+ \leftarrow \operatorname{argmin}_{U'} \tilde{F}^\lambda_{t,U_k}(U')$
end while
Output: t, U_+

To discuss the complexity of the algorithm, we assume that $X^T X$ and $Y^T X$ are computed in advance. Although the computation of z_U requires an SVD of

$Y^T XU$, the latter costs only $O(kr^2)$. Computing $\nabla f(U)$ has then a complexity of $O(p^2 r + pkr)$. The biconvex formulation of Park et al. [2016] leads to iterations with the same theoretical complexity for RRR but it is incompatible with SRRR. Additionally, experiments show that our algorithm is faster (*cf.* Section 5.6 and Appendix G.13).

5.4 Global convergence results

Although recent papers such as Lee et al. [2017] have shown that the gradient descent algorithm escapes saddle points by leveraging the strict saddle property, global convergence for Algorithm 1 is not obvious because f is not smooth. Besides, to the best of our knowledge, none of the papers that exclude convergence towards saddle points deals with regularizers or line search.

5.4.1 Convergence to a critical point for RRR

For RRR, results of Khamaru and Wainwright [2018] apply to our formulation and show that our algorithm converges towards a critical point. Precisely, f_1 is continuously differentiable with Lipschitz gradients, f_2 is continuous and convex and the difference f is bounded below by $-\frac{1}{2}\|Y\|_F^2$. Besides, as a difference of semi-algebraic functions, f satisfies the Kurdyka-Łojasiewicz property whose definition is given in Appendix G.2.4. Therefore, for gradient descent, our setting satisfies the conditions of Theorems 1 and 3 of Khamaru and Wainwright [2018] and we can prove that our algorithm converges from any initial point to a critical point in the sense of Definition 29 in Appendix G.2.5. This is more formally stated in Appendix G.6.1.

5.4.2 Convergence to a critical point for SRRR

In addition to the properties of f_1 and f_2 discussed above in Section 5.4.1, the norm $\|\cdot\|_{1,2}$ is clearly proper, lower semi-continuous and convex so our setting for proximal gradient descent satisfies the conditions of the first part of Theorem 2 in Khamaru and Wainwright [2018]. The latter can be adapted to prove that all limit points of the sequence are critical points in the sense of Definition 29 in Appendix G.2.5. However, to prove actual convergence of the sequence, their Theorem 4 formally requires that f_2 is a function with locally Lipschitz gradient, which is not true when $Y^T XU$ is not full-rank.

Actually, an inspection of the proof of Theorem 4 in Khamaru and Wainwright [2018] shows that the local smoothness condition is only required in a neighborhood of the limit points of the sequence. We prove in Appendix G.6.2 that if all groups of at least r rows of $X^T Y$ are assumed full-rank, which holds almost surely if X and Y contain for example continuous additive noise, and unless local minima are so sparse that the number of selected variables is strictly smaller than r, then any local minimum $U \in \mathbb{R}^{p,r}$ is such that $Y^T XU$ is full-rank. As a consequence, if we assume that the limit points of the sequence produced by the algorithm are a subset of the local minima, then these limit points are contained within a compact set where the

function is smooth and the proof of Theorem 4 of Khamaru and Wainwright [2018] can be adapted in a straightforward manner to obtain global convergence.

5.5 Local convergence analysis

In this section, we prove linear convergence rates in a neighborhood of the global minima for RRR and under a condition on the regularization parameter λ for SRRR. Precisely, we first study the geometry around the optima of (RRR) via a change of variables. Then, a continuity argument shows that the structure remains approximately the same for (SRRR) with a small $\lambda > 0$. Finally, we introduce and leverage Polyak-Łojasiewicz inequalities to prove local linear convergence.

5.5.1 A key reparameterization for RRR

The relation between RRR and PCA and the form of the analytical solution given by Velu and Reinsel [2013] will allow us to show that our study of the objective of RRR can be reduced to the study of the particular case in which X and Y are full-rank diagonal matrices, via a linear change of variables based on the singular value decomposition PSQ^T introduced in Section 5.3.2 of the matrix $(X^TX)^{-\frac{1}{2}}X^TY$. From now on, we assume that the rank parameter r is smaller than the rank of X^TY *i.e.* $r \leq \ell := \operatorname{rank}(X^TY)$. It makes sense to assume that the imposed rank is less than the rank of the optimum for the unconstrained problem, otherwise the rank constraint is essentially useless. We also assume[1] that $s_1 > \ldots > s_\ell$ and that X^TX is invertible.

With the notations of Section 5.3.2, let $P^\perp \in \mathbb{R}^{p,p-\ell}$ be a matrix such that ${P^\perp}^T P^\perp = I_{p-\ell}$ and $P^T P^\perp = 0$, and consider the linear transformation $U = \tau(A, C)$ where :

$$\tau : \begin{cases} \mathbb{R}^{\ell,r} \times \mathbb{R}^{p-\ell,r} \to \mathbb{R}^{p,r} \\ (A, C) \mapsto (X^TX)^{-\frac{1}{2}}(PSA + P^\perp C) \end{cases}. \tag{5.6}$$

Defining $f_a(A) = \frac{1}{2}\|SA\|_F^2 - \|S^2A\|_*$, we show in Appendix G.7.1 that :

$$(f \circ \tau)(A, C) = f_a(A) + \frac{1}{2}\|C\|_F^2 . \tag{5.7}$$

Since τ is invertible, the minimization in (RRR) w.r.t. U is equivalent to the minimization of $f \circ \tau$ w.r.t. (A, C). We can therefore study the original optimization problem by studying f_a.

Similarly to Baldi and Hornik [1989], we characterize the minima of f_a using the connection between PCA and RRR, with a proof given in Appendix G.7.2.

Lemma 9. *The set of minima of f_a is :*

$$\Omega_a^* := \left\{\tilde{I}R \mid R \in \mathcal{O}_r\right\} \quad \textit{with} \quad \tilde{I} := \begin{bmatrix} I_r \\ 0_{\ell-r,r} \end{bmatrix} \in \mathbb{R}^{\ell,r}.$$

[1] These assumptions are also reasonable and will hold in particular if (X, Y) are drawn from a continuous density. We discuss the case where X^TX is not invertible in Appendix G.7 and in Appendix G.8.2, we show why these assumptions are needed.

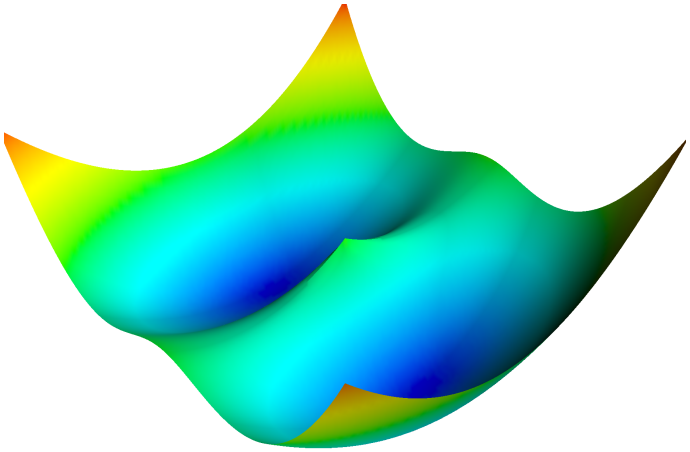


FIGURE 5.1: Graph of f_a for $A \in \mathbb{R}^{2,1}$. In this particular case, $\Omega_a^* = \{(1;0),(-1;0)\}$ and $\mathcal{O}_1 = \{-1,1\}$.

In words, Ω_a^* is the image by the linear transformation $R \mapsto \tilde{I}R$ of the Stiefel manifold $\mathcal{O}_r := \left\{R \in \mathbb{R}^{r,r}, R^T R = I_r\right\}$. In particular, Ω_a^* has two connected components. We also classify the critical points of f_a in Appendix G.7.3 :

Lemma 10. *Rank-deficient matrices cannot be critical points of f_a. Critical points of f_a among full-rank matrices are differentiable points and either global minima or saddle points. Therefore, all local minima of f_a are global.*

5.5.2 Local strong convexity on cones

Although f_a is not convex even in the neighborhood of its minima, we will show that it is locally convex around them in the subspace orthogonal to the set of minima. For any $A \in \mathbb{R}^{p,r}$, let :

$$\Pi_{\Omega_a^*}(A) := \operatorname*{argmin}_{B \in \Omega_a^*} \|B - A\|_F^2$$

be the closest minima to A, and for any $R \in \mathcal{O}_r$, let $\mathcal{C}_a(R)$ be defined as follows :

$$\mathcal{C}_a(R) := \{A \in \mathbb{R}^{\ell,r} \mid \tilde{I}R \in \Pi_{\Omega_a^*}(A)\}.$$

$\mathcal{C}_a(R)$ is the set of points that are associated with the same minimum parameterized by $\tilde{I}R$. As shown in the following lemma, the sets $\mathcal{C}_a(R)$ are actually convex cones that are images of each other by orthogonal matrices; this result is essentially a consequence of the polar decomposition and of the orthogonal invariance of f_a. Let $\mathcal{S}_r^+ \subset \mathbb{R}^{r,r}$ denote the set of positive-semidefinite matrices.

Lemma 11. *For each $R \in \mathcal{O}_r$, $\mathcal{C}_a(R)$ is a cone in $\mathbb{R}^{\ell,r}$ and :*

$$\mathcal{C}_a(I_r) = \left\{ \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \mid A_1 \in \mathcal{S}_r^+,\ A_2 \in \mathbb{R}^{\ell-r,r} \right\}, \tag{5.8}$$

$$\mathcal{C}_a(R) = \{AR \mid A \in \mathcal{C}_a(I_r)\} \quad \textit{and} \quad \bigcup_{R \in \mathcal{O}_r} \mathcal{C}_a(R) = \mathbb{R}^{\ell,r}.$$

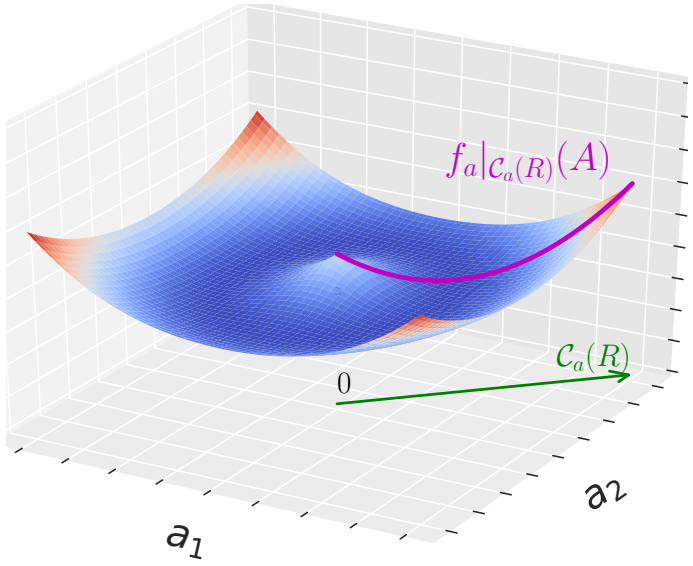


FIGURE 5.2: Schematic 2D graph of f_a around one of the connected components of Ω_a^* when $r \geq 2$. Here, the component of Ω_a^* is a circle and the cones are half-lines with extreme points at the origin.

Note that the cones $\mathcal{C}_a(R)$ do not form a partition of $\mathbb{R}^{\ell,r}$ because if A_1 is not invertible, its polar decomposition is not unique so $[A_1^T \ A_2^T]^T$ is in several cones. However the relative interiors of all the cones partition the set of matrices $[A_1^T \ A_2^T]^T$ such that A_1 is invertible (*cf.* Fact 59 in Appendix G.8.1). The decomposition on these cones is motivated by the fact that for $r \geq 2$, the function f_a in a neighborhood of each of the two connected components of Ω_a^* can be informally thought of as having the shape of the base of a glass bottle with a punt. This is illustrated in Figure 5.2.

Thus, given $R \in \mathbb{R}^{r,r}$, we focus on the restriction $f_a|_{\mathcal{C}_a(R)}$ of f_a on the cone $\mathcal{C}_a(R)$. The next result states in particular that $f_a|_{\mathcal{C}_a(R)}$ is smooth and strongly convex[2] in a neighborhood of $\tilde{I}R$.

Theorem 12. *For any $0 < \mu_a < s_\ell^2(1 - \frac{s_r^2}{s_{r+1}^2})$, there exists a non-empty sublevel set $\mathcal{V}_a \subset \mathbb{R}^{\ell,r}$ of f_a such that f_a is s_1^2-smooth in $\mathcal{V}_a$ and for any $R \in \mathcal{O}_r$, the restriction $f_a|_{\mathcal{C}_a(R)}$ is μ_a-strongly convex in $\mathcal{V}_a \cap \mathcal{C}_a(R)$.*

Via τ these properties of f_a carry over to f. Let ν_X and L_X be respectively the smallest and largest eigenvalues of X^TX and $\mathcal{C}(R) := \tau(\mathcal{C}_a(R), \mathbb{R}^{p-\ell,r})$ with τ defined in Equation (5.6).

Corollary 13. *For any $0 < \mu < \nu_X(1 - \frac{s_{r+1}^2}{s_r^2})$, there exists a non-empty sublevel set $\mathcal{V}^0$ of the function f that can be partitioned into disjoint convex elements $\{\mathcal{C}(R) \cap \mathcal{V}^0\}_{R \in \mathcal{O}_r}$ such that f is L_X-smooth on $\mathcal{V}^0$ and is μ-strongly convex on every $\mathcal{V}^0 \cap \mathcal{C}(R)$.*

To extend partially the previous result to (SRRR), we apply Theorem 6.4 of Bonnans and Shapiro [1998] : given that (a) the objective F^λ of (SRRR) is locally

[2]The definitions of μ-strong convexity, L-smoothness and sublevel sets are recalled in Appendix G.2.

strongly convex on the cone $\mathcal{C}(I_r)$ around the minimum, (b) for every fixed λ in some interval $[0, \tilde{\lambda})$, f is locally Lipschitz with a constant that does not depend on λ and, (c) $F^\lambda - F^0 = \lambda \|\cdot\|_{1,2}$ is locally Lipschitz with a constant $\sqrt{p}\lambda$ which is $O(\lambda)$, then by Bonnans and Shapiro [1998, Theorem 6.4], there exists $\check{\lambda} > 0$ such that for all $0 \leq \lambda < \check{\lambda}$, the minimum of F^λ in $\mathcal{C}(I_r)$ is a continuous function of λ. This is detailed in Appendix G.8.4.

Corollary 14. *There exists $\bar{\lambda} > 0$ such that for any $0 \leq \lambda < \bar{\lambda}$ and $0 \leq \mu < \nu_X(1 - \frac{s_{r+1}^2}{s_r^2})$, there exists a non-empty sublevel set $\mathcal{V}^\lambda$ of F^λ that can be partitioned into disjoint convex elements $\{\mathcal{C}(R) \cap \mathcal{V}^\lambda\}_{R \in \mathcal{O}_r}$ so that f is L_X-smooth on $\mathcal{V}^\lambda$ and F^λ is μ-strongly convex on every $\mathcal{C}(R) \cap \mathcal{V}^\lambda$.*

These characterizations of the geometry in a neighborhood of the optima immediately lead to Polyak-Łojasiewicz inequalities that entail the linear convergence of first-order algorithms.

5.5.3 P-Ł inequalities and proofs for linear convergence rates

Polyak-Łojasiewicz (**PŁ**) and Kurdyka-Łojasiewicz inequalities (**KŁ**) were introduced to generalize to nonconvex functions (or just not strongly convex) proofs of rates of convergence for first-order methods [Attouch and Bolte, 2009; Karimi et al., 2016, and references therein]. In particular, PŁ generalizes the fact that, for a differentiable and μ-strongly convex function f with optimal value f^*,

$$f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2 . \tag{PŁ}$$

Karimi et al. [2016] and Csiba and Richtarik [2017] proposed a generalization to a *proximal* PŁ inequality of relevance for forward-backward algorithms applied to non-differentiable functions. In this section , we summarize an immediate extension allowing a line search procedure, of results established for first-order algorithms to prove locally a linear rate of convergence. Consider $d \in \mathbb{N}^*$ and a function[3] $F^\lambda = f + \lambda h$ defined on $\mathbb{R}^d$ and with optimal value $F^{\lambda,*}$, where f is an L-smooth function and h is a proper lower semi-continuous convex function. We define the t-approximation $\tilde{f}_{t,x}$ and $\tilde{F}^\lambda_{t,x}$ of f and F^λ at x as in Section 5.3.3. The t-decrease function is defined as :

$$\gamma_t(x) := -\frac{1}{t} \min_{x' \in \mathbb{R}^d} \left[\tilde{F}^\lambda_{t,x}(x') - F^\lambda(x) \right] . \tag{5.9}$$

Given x, assume that the minimum in Equation (5.9) is attained at a point x^+ for $t > 0$ such that the (LS) condition $\tilde{F}^\lambda_{t,x}(x^+) \geq F^\lambda(x^+)$ is satisfied. Then the decrease in the objective value $F^\lambda(x) - F^\lambda(x^+)$ is lower bounded by $t\gamma_t(x)$, hence the name t-decrease function (see Fact 41 in Appendix G.5.1). We make use of a natural generalization of the *proximal* PŁ inequality proposed by Karimi et al. [2016] and Csiba and Richtarik [2017]. For x such that $F^\lambda(x) > F^{\lambda,*}$, with $F^{\lambda,*}$ the minimum of F^λ, we define the t-proximal forcing function :

$$\alpha_t(x) := \frac{\gamma_t(x)}{F^\lambda(x) - F^{\lambda,*}}.$$

[3] In this section we use a general variable x but we keep using f and F^λ.

We can now state the following theorem that bounds the optimal gap for our algorithm iteratively.

Theorem 15. *[From Lemma 13 in Csiba and Richtarik, 2017] Let $x \in \mathbb{R}^d$ and x^+ be defined by $x^+ = \text{argmin}_{x'}[\tilde{F}^{\lambda}_{t,x}(x') - F^{\lambda}(x)]$, where t is chosen so that the line search condition (LS) is satisfied. Then we have :*

$$F^{\lambda}(x^+) - F^{\lambda,*} \leq [1 - t\,\alpha_t(x)]\,\left[F^{\lambda}(x) - F^{\lambda,*}\right].$$

Given $t > 0$, we say that F^{λ} satisfies the (t-strong *proximal* PŁ) inequality in a set $\mathcal{V} \subset \mathbb{R}^d$ if there exists $\alpha(t) > 0$ such that for any $x \in \mathcal{V}$ where $F^{\lambda}(x) > F^{\lambda,*}$, we have :

$$\alpha_t(x) \geq \alpha(t). \qquad \text{(t-strong } \textit{proximal} \text{ PŁ)}$$

If $\lambda h = 0$, then $\gamma_t(x) = \frac{1}{2}\|\nabla f(x)\|^2$ and it is easy to see that (t-strong *proximal* PŁ) boils down to (PŁ) with $\mu = \alpha(t)$.

5.5.4 Proving local linear convergence

We now return to the functions f and F^{λ} defined for (RRR) and (SRRR) with minimal values f^* and $F^{\lambda,*}$, and we establish the (PŁ) and (t-strong *proximal* PŁ) inequalities in a neighborhood of their respective global minima.

Corollary 16. *Let $0 \leq \mu < \nu_X(1 - \frac{s^2_{r+1}}{s^2_r})$ and $\mathcal{V}^0$ as in Corollary 13. For all $U \in \mathcal{V}^0$, we have :*

$$f(U) - f^* \leq \frac{1}{2\mu}\|\nabla f(U)\|_F^2.$$

In light of Corollary 14, we can also prove the (t-strong *proximal* PŁ) inequality for F^{λ} with small values of λ. To this end, we consider $\bar{\lambda} > 0$ as in Corollary 14.

Corollary 17. *Let $0 \leq \mu < \nu_X(1 - \frac{s^2_{r+1}}{s^2_r})$ and $0 \leq \lambda < \bar{\lambda}$. For any $t > 0$, F^{λ} satisfies the (t-strong proximal PŁ) inequality with $\alpha(t) := \min(\frac{1}{2t}, \mu)$. In other words, for any $t > 0$ and $U \in \mathcal{V}^{\lambda}$, we have :*

$$\gamma_t(U) \geq \alpha(t)\,\left[F^{\lambda}(U) - F^{\lambda,*}\right],$$

$$\textit{with } \gamma_t(U) := -\frac{1}{t}\min_{U' \in \mathbb{R}^{p,r}}\left[\tilde{F}^{\lambda}_{t,U}(U') - F^{\lambda}(U)\right].$$

So, leveraging Theorem 15 and Corollary 16/17 for (RRR)/(SRRR), we obtain the linear rate of convergence which is proved in Appendix G.10.3. Indeed, if L_X denotes the largest eigenvalue of X^TX and β the step-size decrease factor in Algorithm 2, then we have the following result.

Corollary 18. *Let $0 \leq \lambda < \bar{\lambda}$ and $k \geq 0$. Assume that $t_{k-1} > \frac{\beta}{L_X}$ and U_{k+1} is generated as in Algorithm 1 from $U_k \in \mathcal{V}^{\lambda}$. Then $U_{k+1} \in \mathcal{V}^{\lambda}$, $t_k > \frac{\beta}{L_X}$ and denoting $\rho = \min(\frac{1}{2}, \beta\frac{\mu}{L_X})$, we have :*

$$F^{\lambda}(U_{k+1}) - F^{\lambda,*} \leq (1 - \rho)\,\left[F^{\lambda}(U_k) - F^{\lambda,*}\right].$$

As explained in Fact 43 in Appendix G.5.2, there is only a finite number of steps at the beginning of Algorithm 1 for which the assumption $t_k > \frac{\beta}{L_X}$ may fail. The convergence is therefore linear. We propose a direct proof of Corollary 18 based on Corollary 17 and Theorem 15. It should be noted that the geometric structure leveraged for Corollary 17 can also be used to obtain Corollary 18 as a consequence of the Kurdyka-Łojasiewicz inequality (*cf.* Appendix L).

5.6 Experiments on RRR and SRRR

We perform experiments on simulated data both for RRR and SRRR to assess the performance of the algorithms in terms of speed.

For RRR, we compare gradient descent algorithms in U space and in (U, V) space. In the former case, we just minimize (RRR), whereas in the latter, following Park et al. [2016], we minimize $\mathcal{F}(UV^\top) + g(U, V)$, with $\mathcal{F}(W) = \frac{1}{2}\|Y - XW\|_F^2$ and $g(U, V) = \frac{1}{4}\|U^\top U - V^\top V\|_F^2$; this objective has the same optimal value as $\mathcal{F}(UV^\top)$, but the regularizer g was shown to improve the convergence rate of the algorithm (see Appendix G.13.1). Note that the formulation of Park et al. [2016] does not apply to SRRR because the regularizer g is not compatible with the use of the group-Lasso norm.

For SRRR, we implement proximal gradient descent algorithms and compare in speed with the RRR case and with the alternating optimization algorithm proposed[4] in Bunea et al. [2012]. In each case we consider variants of these first-order methods with and without line search.

For the alternated procedure, each inner minimization of the matrix U is stopped when a duality gap becomes smaller than the desired precision 10^{-4}. Since it takes more than seconds to optimize, it justifies the relevance of RRR/SRRR.

As in Bunea et al. [2012], we sample the rows of X from a zero-mean Gaussian with a Toeplitz covariance matrix Σ where $\Sigma_{i,j} = \rho^{|i-j|}$ and $\rho \in (0, 1)$. We set $n = 10^3$, $p = 300$ and $k = 200$. We let $W_0 = U_0 V_0^\top$ for $U_0 \in \mathbb{R}^{p,r}$ and $V_0 \in \mathbb{R}^{k,r}$ uniformly drawn from the set of orthonormal matrices with $r_0 = 30$ columns. For SRRR, each row of W_0 is then set to zero with probability p_0. Then we compute $Y = XW_0 + E$ for E a matrix of i.i.d. centered scalar Gaussians with standard deviation $\sigma = 0.1$. Finally, we solve all formulations for a matrix W of rank $r = 20$. For all algorithms, we initialize U (and V if relevant) at random.

We report results for $\rho = 0.6$ in Figure 5.3 and in Appendix G.13 for additional values of ρ and p_0. For RRR, these results show that the algorithms based on our proposed formulation are significantly faster, both in terms of the number of function/gradient evaluations and in terms of time; moreover they benefit more from the line search. We do not report curves with both line search and acceleration because this combination does not yield any speed increase.

For (SRRR) and (RRR) all algorithms exhibit at least linear convergence. Compared with (RRR), the convergence for (SRRR) typically seems as fast or faster. Additionally, the line search plays a significant role in accelerating the convergence of the algorithm.

[4] Ma and Sun [2014] consider a similar algorithm.

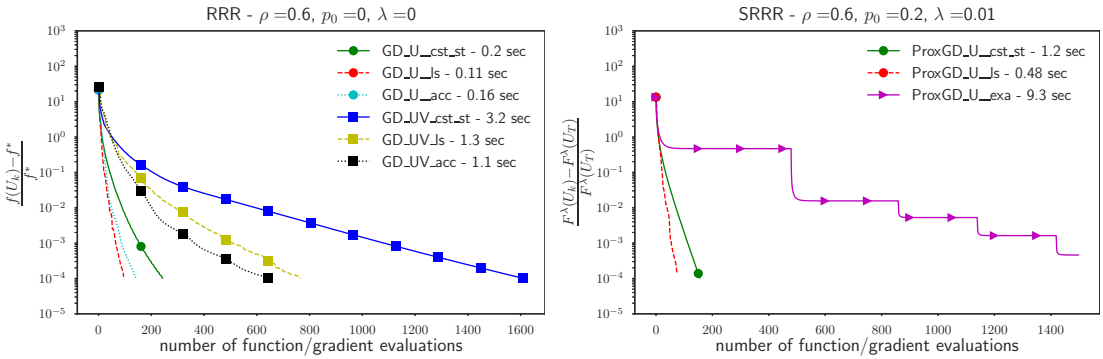


FIGURE 5.3: *(Left)* RRR : Convergence of $f(U_k) - f^*$ for gradient descent on our formulation in U with constant step size (`GD_U_cst_st`), with line search (`GD_U_ls`), with the acceleration (`GD_U_acc`) proposed by Li and Lin [2015] and gradient descent for the formulation of [Park et al., 2016] with constant step size, line search and acceleration (`GD_UV_cst_st`, `GD_UV_ls`, `GD_UV_acc`). *(Right)* SRRR with $\lambda = 0.01$: Convergence for T large of $F^\lambda(U_k) - F^\lambda(U_T)$ for proximal gradient descent on our formulation with and without line search (`ProxGD_U__ls`, `ProxGD_U__cst_st`), compared with the alternating optimization algorithm (`ProxGD_U__exa`) proposed in Bunea et al. [2012]. The running time to reach a precision of 10^{-4} is given at the top right.

Conclusion of Chapter 5

We considered a reformulation of RRR and SRRR problems as non-convex and non-differentiable optimization problems w.r.t. to a matrix U with r columns. We proposed to apply subgradient-type algorithms studied by Khamaru and Wainwright [2018], which correspond essentially to gradient descent for RRR and proximal gradient descent for SRRR.

The algorithms are provably convergent to critical points under reasonable assumptions. We show that for a certain range of regularization coefficients λ the objective satisfies a Polyak-Łojasiewicz inequality in a neighborhood of the global minima, which entails local linear convergence if the algorithm converges to them.

For RRR, gradient descent converges to a critical point and if a global minimum of the original objective has been found, it can easily be certified.

Future work could try to determine if convergence to saddle points of SRRR can be excluded and if global linear convergence results can be obtained. Another interesting direction of research is to extend these types of results to other matrix optimization problems with low-rank constraints.

Chapter 6

Conclusion of the manuscript

Description of the load forecasting problems We have proposed in Chapter 2 an introduction to the load forecasting problems and a preliminary exploration of the database. Thus, we have had the opportunity to describe the 3 common cycles of the electricity demand, namely the daily, weekly and yearly cycles, along with the conditional expectations of the load with respect to the calendar information and the meteorological conditions. We also described the different settings that we consider and insisted on the variability, as well as the irrelevant values encountered in the load measurements at the level of the substations. Chapter 2 also allowed us to present the detection and correction procedures that we have used to clean the database.

A standard bivariate linear model In Chapter 3, we have described a transformation of the inputs to feed thereafter a linear model tuned by minimizing a classical measure of the squared errors. The same model is proposed to model the load both at the national level and the local levels, with adapted regularization hyperparameters. Unlike the state-of-the-art GAM that include different submodels for the different hours of the day, we have introduced a single modeling, leading to a reasonable computational time with performances comparable to state-of-the-art results and to be used at all times of the day and the year. It is consequently simpler to analyze. The observation of the results with this model allowed us in particular to underline the importance of the interactions between some of the inputs.

Modeling difficulties Meanwhile, we illustrated in Chapter 3 the main difficulties that we encountered. These difficulties are accentuated at the level of the substations because of the higher variability of the load curves. For instance, non-stationarity has been emphasized and we have pointed several works to alter the resulting problems. We have also identified Mondays in particular as a difficult time period to forecast. We consider that this is due to their following Sundays and the fact that we use in the modeling the delayed loads in a rather basic manner. Although we have considered simple patches, like the interaction between the past loads and the hour of the week, we believe that a dedicated modeling would be worth studying. In particular, we have not considered so far the possibility of having different hyperparameters for the different days of the week.

Implementation We have tried to propose an algorithmic framework simultaneously for the national and the local load forecasting problems and we have not insisted in the manuscript on the problems related to the implementation of the models. Nevertheless, the dimensions of the optimization problems associated to the different levels of aggregation are significantly different. It was ambitious to use a single tool for the different aggregation levels because the most efficient algorithmic tools depend on the size of the problems. Thus, we believe that while we encompassed the different problems in a single algorithm, some decisions are suboptimal and it might be relevant to consider a specific algorithm for each problem separately from the others.

Similarity structure In the end of Chapter 3 and in the beginning of Chapter 4, we illustrated the similarities between the independent models learned for each substation in order to motivate the multi-task framework discussed in Chapter 4. We consider that the proposed illustrations are not entirely satisfying so far : how to measure quantitatively the similarities between different models with distinct inputs remains an open question. In particular, we could not convincingly decide which substations should be coupled and which ones should be isolated, if relevant. Still, the presented figures illustrate the presence of a common structure in the learned coefficients for the different substations, that is the task structure, and in the residuals of the models, which corresponds to the output structure.

Task structure The clustering models and in particular the low-rank models, both leveraging the task structure, allowed us to conclude that the number of parameters in the independent models of Chapter 3 is unnecessarily large. The results with the different variants of these models also point out that even if some parameters are shared by the models, a sufficient flexibility is necessary to obtain results comparable with the state-of-the-art models. We believe that it is worthwhile pursuing the research of flexible multi-task models, mixing shared and individual components.

Sparse Reduced Rank Regression The interest for the low-rank constraint motivated the analysis in Chapter 5 of Sparse Reduced Rank Regression, which is a non-convex and non-differentiable optimization problem with respect to a thin matrix U. In particular, we proved the convergence to critical points under reasonable assumptions of a subgradient-type algorithm, which correspond essentially to a proximal gradient descent. We also proved local linear convergence for a certain range of regularization coefficients leveraging a Polyak-Łojasiewicz inequality satisfied by the objective in a neighborhood of the global minima.

Output structure In the last section of Chapter 4, we have considered the output structure and the possibility of coupling the models at different scales. While we have not found a procedure to screen information sharing, we have tried to identify the relevant level to couple the models. We obtained positive results by ensuring that the forecasts are consistent at the districts levels, thereby improving the accuracy of the local models in some districts.

Selectively screen the sharing of information As a conclusion, the models and the results in Section 4.6 support the interest of a multi-task approach. They provide a guarantee for the TSO of having reasonable forecasts both at local and aggregated levels. Although we spent a significant time trying to couple the 1751 models of all the substations, the empirical results also indicate that coupling at a smaller scale is not only less demanding computationally speaking, it also seems more relevant. Eventually, we consider that the research of the most relevant levels for coupling the local models and the development of a procedure to screen information sharing are the next priorities. The analysis of the estimated clusters in Section 4.4 and the low-rank matrices in Section 4.5 is a potential way to illustrate the underlying structure and screen information sharing.

Appendix A

Notations

Sets of numbers

- The set of natural non-negative integers is denoted $\mathbb{N}$, from which we define $\mathbb{N}^* := \mathbb{N}\backslash\{0\}$ and $\mathbb{N}\backslash\{0, 1\}$.
- The set of natural integers is denoted $\mathbb{Z}$, from which we define $\mathbb{Z}^*$.
- The set of real numbers is denoted $\mathbb{R}$, the non-zero real numbers $\mathbb{R}^*$ and the non-negative real numbers $\mathbb{R}_+$.
- The interval between two numbers $x \leq y$ is denoted $[x, y]$.
- The set of integers between $p \in \mathbb{Z}$ and $q \in \mathbb{Z}$ with $p \leq q$ is denoted $[\![p, q]\!]$.
- The set of equivalence classes of numbers modulo 1 is denoted with the torus $\mathbb{R}/\mathbb{Z}$.

Variables

- Scalar observations and coefficients are written with a normal font *e.g.* b, x, y.
- Random scalar variables are written with a sans-serif font *e.g.* x, y.
- Vector of observations and coefficients are written in bold *e.g.* $\boldsymbol{b}, \boldsymbol{x}, \boldsymbol{y}$.
- Random vectors are written with a sans-serif font in bold *e.g.* $\mathbf{x}, \mathbf{y}$.
- The i-th element of a vector $\boldsymbol{b}$ is denoted b_i, unless explicitly stated otherwise.
- Matrices of observations and coefficients are capitalized and bold *e.g.* $\boldsymbol{B}, \boldsymbol{X}, \boldsymbol{Y}$.
- The i-th row of a matrix $\boldsymbol{B}$ is denoted $\boldsymbol{b}_i$, unless explicitly stated otherwise.
- The j-th column of a matrix $\boldsymbol{B}$ is denoted $\boldsymbol{b}^{(j)}$.
- The element in the i-th row and j-th column of a matrix $\boldsymbol{B}$ is denoted b_i^j.
- Tensors of observations and coefficients are capitalized *e.g.* $\boldsymbol{\mathcal{X}}$.
- Given two vectors $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^q$, the element in the i-th row and j-th column of the matrix $\boldsymbol{a} \otimes \boldsymbol{b} \in \mathbb{R}^{p,q}$ is $a_i b_j$.
- Given $\boldsymbol{s} \in \mathbb{R}^p$, $\mathrm{diag}(s_1, \ldots, s_p) \in \mathbb{R}^{p,p}$ is a diagonal matrix with elements $s_1, \ldots, s_p$ on the diagonal.
- Given a vector $\boldsymbol{\ell} \in \mathbb{R}^n$, the average of its elements is denoted $\bar{\ell}$.
- Given $K \in \mathbb{N}^*$, the constant vector denoted $\mathbf{1}_K$ contains only 1s.

Power sets

- The set of real-valued vector of size p is denoted $\mathbb{R}^p$.
- The set of real-valued matrices of size (p, q) is denoted $\mathbb{R}^{p,q}$.
- The set of real-valued tensors of size (p, q, r) is denoted $\mathbb{R}^{p,q,r}$.
- Given two sets A and B, the set of functions from A to B is denoted B^A, thus the set of real-valued functions defined in $\mathbb{R}$ is denoted $\mathbb{R}^{\mathbb{R}}$.
- A sequence of $p \in \mathbb{N}^*$ real-valued functions defined in $\mathbb{R}$ is denoted $(\mathbb{R}^{\mathbb{R}})^p$.
- An array of $p \times q$ real-valued functions defined in $\mathbb{R}$ is denoted $(\mathbb{R}^{\mathbb{R}})^{p,q}$.
- An array with size (a, b) of elements included in $\{0, 1\}$ is denoted $\{0, 1\}^{a,b}$.

Attributes

- The rank of a matrix $\boldsymbol{M}$ is denoted $\text{rank}(\boldsymbol{M})$.
- The transpose of a matrix $\boldsymbol{M}$ is denoted $\boldsymbol{M}^T$.
- The cardinal of a set S is denoted $|S|$.
- The positive part of a number x is denoted $(x)_+ := \max(x, 0)$.
- The factorial of a non-negative number $n \in \mathbb{N}$ is denoted $n!$.
- The binomial coefficient indexed by $k \leq n$ is denoted $\binom{n}{k}$.

Norms

- The 2-norm of a vector $\boldsymbol{b}$ is denoted $\|\boldsymbol{b}\|_2$.
- The Frobenius norm of a matrix $\boldsymbol{M}$ is denoted $\|\boldsymbol{M}\|_F$.
- The Frobenius scalar product between vector or matrices is denoted $\langle \cdot, \cdot \rangle$.
- The trace-norm of a matrix $\boldsymbol{M}$ is denoted $\|\boldsymbol{M}\|_*$, it is the sum of its singular values.

Appendix B

Data cleansing procedure

B.1 Detection of anomalies

In this section, we highlight the fact that the database contains a significant number of irrelevant values. We also present *ad hoc* tools to detect these anomalies. These irrelevant values can be due to errors in measurements, errors in the correction procedure of Section 2.1 or modifications of the network configuration.

Errors in measurements There are three types of errors in the database that are particularly easy to detect : the *Not a number* values, the negative values and the zero or very close to zero values. The *Not a number* values and the zero values probably correspond to errors in measurements while the negative values may follow from an overestimation of the local renewable production in the procedure of Section 2.1. The distribution of the number of anomalous values is presented in Figure B.1.

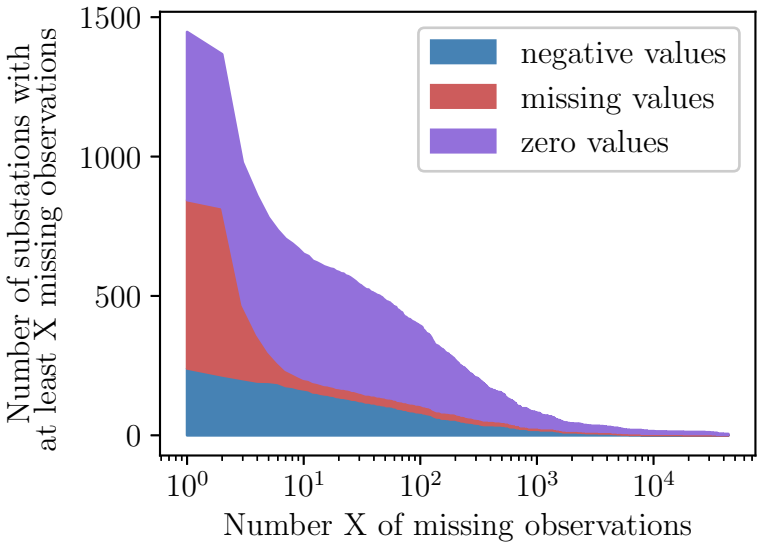


FIGURE B.1: Number of anomalous values in the database
Repartition of the anomalous values among the substations. For instance, 200 substations have at least 300 anomalous values.

Load reports and anomalies There are in the database inconsistencies that are more difficult to detect. Load reports for instance, correspond to the transfer of a fraction of the load of one substation on another substation. This mechanism leads to load curves like in Figure B.2. There are other anomalies in the database that we cannot, for sure, attribute to load reports. However, the tools that we use to detect them are the same.

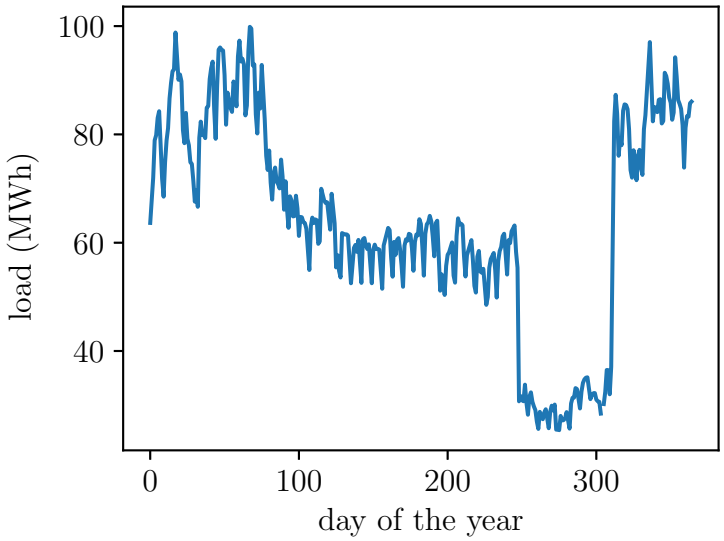


FIGURE B.2: Illustration of a load report
Average load per day over one year to illustrate a load report at one of the substation in the database. A fraction of the load of the substation is reported on other substations from August to November, thus leading to the jumps and the decrease of the load during this period.

Detection with trimmed means In order to detect anomalies in the database, we first use trimmed means. Given an observation instant i and a measurement of the load ℓ_i at a substation whose mean is denoted $\bar{\ell}$, we extract from the database the loads $\{\ell_{i+24j}\}_{j=-14,\dots,14}$ at the same hour of the day during the preceding two weeks and the following two weeks. From this set, we remove the maximum and minimum values and compute the mean μ_i of the remaining samples. Given a threshold $\tau = \frac{\bar{\ell}}{10}$, the observation instant i is classified as an error if $|\ell_i - \mu_i| > \tau$. The choice of the threshold and the 1 month long window have not been optimized.

Detection with middle-term models Another trick to detect anomalies relies on the observation of the residuals with a load forecasting model that is only based on calendar and weather information and does not include the recent loads. Such models are called middle-term models and are detailed in Section 2.5.1.

Empirically, we observed that large residuals at a given observation instant usually correspond to a jump of the load time series. This procedure is not automatic but still allowed us to manually identify substations with irregularities.

Altogether, about 800 substations present notable anomalies. A large part of them are corrected as explained in Section B.2

B.2 Correction of anomalous values

After detecting the irrelevant values with the procedure described in Section B.1, we consider two possibilities : either modifying these values to make the load curves more consistent or squarely remove the concerned substation from the database.

To propose a correction, when an irrelevant value is detected, we could resort to the trimmed mean presented in Section B.1. However, the corrupted data are often consecutive and occur on periods of several days or even weeks, which makes the trimmed mean an irrelevant substitute. Instead, we take advantage of the following observations : Given a set $\mathcal{K}_0$ of substations in the database where no irrelevant value was detected, and a substation κ^* with an irrelevant value at the observation instant $i \in \mathbb{N}$, a remarkably accurate way to forecast the load $\ell_i^{\kappa^*}$ at the substation κ^* and instant i, is to regress it on the loads $(\ell_i^{\kappa})_{\kappa=1,\dots,\mathcal{K},\kappa\neq\kappa^*}$ at the other substations and the same instant i. Of course, this method cannot be applied for load forecasting because it requires the oracles $(\ell_i^{\kappa})_{\kappa=1,\dots,\mathcal{K},\kappa\neq\kappa^*}$ but we can use it to correct the irrelevant values in the database, with a model estimated on a different time period.

In practice, we choose indeed for $\mathcal{K}_0$ the set of substations where no irrelevant value was detected, there are about 1200 such substations and, given κ^* a substation with irrelevant values at observation instants $\mathcal{I}^*$, we randomly partition the set of sane observation instants instants into 2 subsets $\mathcal{I}^{\text{train}}$ and $\mathcal{I}^{\text{test}}$, respectively containing 80 % and 20 % of the sane observations. Then, we train a regression model with the data in $\mathcal{I}^{\text{train}}$ to predict the load at κ^* with the loads at the substations in $\mathcal{K}_0$ and compute the coefficient of determination (presented in Section 2.7.1) on the test set $\mathcal{I}^{\text{test}}$. Given a threshold $\tau = 0.8$, we keep the substation κ^* in the database if the coefficient of determination on $\mathcal{I}^{\text{test}}$ is above τ and modify the irrelevant values with the trained models for observation instants in $\mathcal{I}^*$. Otherwise, the substation is eliminated from the database.

Obviously, keeping as many substations in the database with consistent values would be ideal but, since our final objective is to study a multi-task forecasting model, the irrelevant values at some substations can represent a significant hindrance for this model. Therefore we allow ourselves, adopting a pragmatical approach, to choose the second option. In practice, 10 to 15 % of the substations are thereby discarded : the resulting database contains 1751 substations.

We do not pretend that these detection and correction mechanisms are optimal but consider that they are sufficient to clean the database from significant errors. We used random forests or regression models with a LASSO penalty for the correction but did not work on the best hyperparameters. This requires further work.

Appendix C

The design matrices

We described in Chapter 3 a general procedure to build a standard bivariate linear model from a set of inputs. In the beginning we have a dataset containing different categories of inputs as distinguished in Table 3.1 and target variables. To describe a procedure as generic as possible, we assumed that all the inputs except for the timestamp lie in the interval $[0, 1]$ after affine transformations. We then proceeded to the feature engineering of Section 3.1. These transformations are described here.

C.1 Restriction to $[0, 1]$

The timestamp The input corresponding to the timestamp is affinely transformed so that the value of the first day in the dataset is 0 and the value of the 365^{th} day is 1. If there are 5 years in the dataset, the maximum value after the affine transformation is 5.

The indicators We used binary indicators for the Christmas period, the holidays, and the hours between the sunrise and the sunset. These inputs are not transformed.

Continuous acyclic inputs The temperatures, the cloud covers and the past loads are continuous acyclic inputs. Originally, they have different scales but they are affinely transformed to lie in the interval $[0, 1]$. For instance, consider a temperature measured at a given substation with extremal values in a training set $T_{\min}$ and $T_{\max}$. Then the following transformation is applied to this input so that it ends in $[0, 1]$:

$$T \mapsto \frac{T - T_{\min}}{T_{\max} - T_{\min}}. \tag{C.1}$$

Similar transformations are used for the past loads and the cloud covers, both for the training sets and the test sets.

Cyclic inputs The cyclic inputs are the day of the year and the hour of the week. For instance, the hour of the week with values in $[0, 167]$ is transformed with

$$x \mapsto \frac{x}{168}. \tag{C.2}$$

The extremal values after this transformation are 0 and $\frac{167}{168}$. We can therefore consider that both ends of the intervals $[0, 1]$ coincide for this cyclic variable. Similarly, the day of the year is transformed with

$$x \mapsto \frac{x}{366}. \tag{C.3}$$

We consider that no specific treatment is needed for leap years as the effect should be marginal.

C.2 Centering and normalization

Feature engineering After the inputs have been affinely transformed as described in Section C.1, we build the covariates with the feature engineering presented in Section 3.1.1 for the univariate features and in Section 3.1.2 for the bivariate features. This procedure leads for each observation to a vector of covariates

$$\mathbf{x} := \left(1, [\boldsymbol{\phi}_d(\xi_d)]_{d\in\mathcal{U}}, [\boldsymbol{\Phi}_{d,e}(\xi_d, \xi_e)]_{(d,e)\in\mathcal{B}}\right) \in \mathbb{R}^p, \tag{C.4}$$

that is obtained by concatenating all the covariates built with the features from the inputs $(\xi_1, \ldots, \xi_D) \in \mathbb{R}^D$.

The design matrix is then obtained by concatenating this covariate vector for the different observations as described in Section 3.3. However, before proceeding to the minimization of the different problems considered in this manuscript, we center and normalize each group of columns by computing for each group of covariates $\boldsymbol{\phi}_d(\xi_d)$ with $d \in \mathcal{U}$ or $\boldsymbol{\Phi}_{d,e}(\xi_d, \xi_e)$ with $(d, e) \in \mathcal{B}$, the variance and the mean of the associated splines. There is only one variance and one mean computed for each $d \in \mathcal{U}$ or $(d, e) \in \mathcal{B}$.

Appendix D

Implementation of GAM benchmarks

In order to ensure the reproducibility of the comparison between our models and the GAM benchmarks in Section 3.5, we detail in this section the implementation that we have used for the GAM. Because the available information is slightly different from the national setting considered by Pierrot and Goude [2011] and the local setting considered by Goude et al. [2013], we provide here the complete formulas used in R with the MGCV library presented in Section 2.9.4.

National GAM The formula used in R to compute the numerical performances of the GAM proposed by Pierrot and Goude [2011] at the national level is :

$$\begin{aligned}\hat{\mathsf{y}} \quad \sim \quad & \sum_{d=1}^{7} 1_{\text{weekday}=d} \\ & + s(\mathsf{T}^0) + s(\mathsf{T}^0_{\text{-}24}) + s(\bar{\mathsf{T}}^0_{\text{-}24}) + s(\underline{\mathsf{T}}^0_{\text{-}24}) + s(\mathsf{T}^0_{\text{-}48}) \\ & + s(\mathsf{c}^0, k=8) + \mathsf{t} + s(\mathsf{d}) + s(\mathsf{d}, by = 1_{\textit{week-end}}) \\ & + \sum_{d=1}^{7} s(\ell_{\text{-}24}, by = 1_{\text{weekday}=d}). \end{aligned} \tag{D.1}$$

We recall that there is one model for each of the 24 hours of the day and we refer the reader to Wood and Wood [2015] for details on these formulas.

Local GAM The formula used to compute the numerical performances of the independent GAM proposed by Goude et al. [2013] at the level of the substations is :

$$\begin{aligned}\hat{\mathsf{y}} \quad \sim \quad & \sum_{d=1}^{7} 1_{\text{weekday}=d} \\ & + \sum_{s=1}^{2} [s(\mathsf{T}^s) + s(\mathsf{T}^s_{\text{-}24}) + s(\mathsf{T}^s_{\text{-}48}) + s(\tilde{\mathsf{T}}^s_{\text{-}99})] \\ & + 1_{\mathsf{hld}} + s(\mathsf{d}) + s(\ell_{\text{-}24}). \end{aligned} \tag{D.2}$$

where $s = 1, 2$ designates the 2 geographically closest weather station to the considered substation and $\tilde{\mathsf{T}}^s_{\text{-}99}$ is the exponential smoothing of the temperature proposed

by Goude et al. [2013]. We recall that there is one model for each of the 24 hours of the day. Again, we refer the reader to Wood and Wood [2015] for details on these formulas.

Appendix E

Implementation details

Except for the GAM models that is based on the MGCV library in *R*, all our models were implemented in *Python* and rely heavily on the *NumPy* Library.

The Scipy.sparse library Large matrices were stored using the sparse formats proposed in the *SciPy* Library. However, no acceleration was observed when performing computations with this library, it was mainly to reduce the use of the Random Access Memory (**RAM**). We believe that the absence of an acceleration is due to the relatively small size of the matrices that we consider and the fact that they are not sparse enough for the offset caused by the use of the sparse library to be neglected.

Speedup with a collective optimization A significant amount of time was devoted to accelerating the computations. For instance, for the minimization of Problem (3.25) where the models for each substation are learned independently, we have observed that it is faster to proceed to a single minimization collectively for all the substations at once, instead of solving sequentially the problems for the different substations.

We believe that this is due to the relatively small dimensions of the matrices that we consider and to the fact that the basic rules to estimate the computational complexity of an algorithm do not apply with such small matrices because the times for reading and writing cannot be neglected. Thereby, we enjoy an acceleration of the computations by collectively solving Problem (3.25) with the tools provided by the *NumPy* library.

This decision raises questions concerning the stopping criteria and when a line search is used, as described in Chapter 5. More precisely, we had to decide wether those quantities should be the same for all the substations. Not having observed significant differences, we do not detail these questions.

Optimization algorithms We have considered two different possible algorithms for the minimization problems presented in this manuscript. In most cases, the problems that we consider are twice continuously differentiable. It is in particular the case when Ridge or Smoothing Splines penalties are used. We have consequently leveraged the speed of the quasi-Newton method of Broyden, Fletcher, Goldfarb, and

Shanno (**L-BFGS**) [Liu and Nocedal, 1989; Zhu et al., 1997] implemented in the *SciPy* library [Jones et al., 2001].

However, we have considered multiple times variable selection procedures, in particular in Chapter 5 with non-differentiable penalties like LASSO or group-LASSO. In this case, we have implemented in *Python* a Block-Coordinate Descent (**BCD**) algorithm with a line search and an active set procedure. The optimization is slower with this first-order descent algorithm.

Appendix F

Additional Figures and Tables

F.1 Additional Figures and Tables for Chapter 2

F.1.1 Weather information

Weather stations	Weights (%)	Weather stations	Weights (%)
Abbeville	1.0	Nantes	4.2
Bale-Mulhouse	2.0	Nevers	1.5
Bordeaux	4.0	Nice	3.6
Boulogne-Sur-Mer	1.0	Nimes-Courbessac	2.4
Bourg-Saint-Maurice	2.75	Orange	1.2
Bourges	4.2	Paris-Montsouris	11.25
Brest-Guipavas	4.2	Paris-Orly	0.0
Caen	2.5	Perpignan	1.6
Clermont-Ferrand	2.75	Reims	0.0
Dijon	1.0	Rennes	4.2
Le Luc	1.2	Saint-Auban	1.2
Lille	3.0	Strasbourg	1.0
Limoges-Bellegarde	3.2	Tarbes-Ossuns	4.0
Lyon-Satolas	5.5	Toulouse-Blagnac	1.6
Marignane	2.4	Tours	4.2
Montpellier	1.6	Trappes	11.25
Nancy-Essey	3.0	Troyes	1.5

TABLE F.1: Weights of the weather stations
Weights associated to the different weather stations for the national load forecasting problem in the historical model [RTE, 2011].

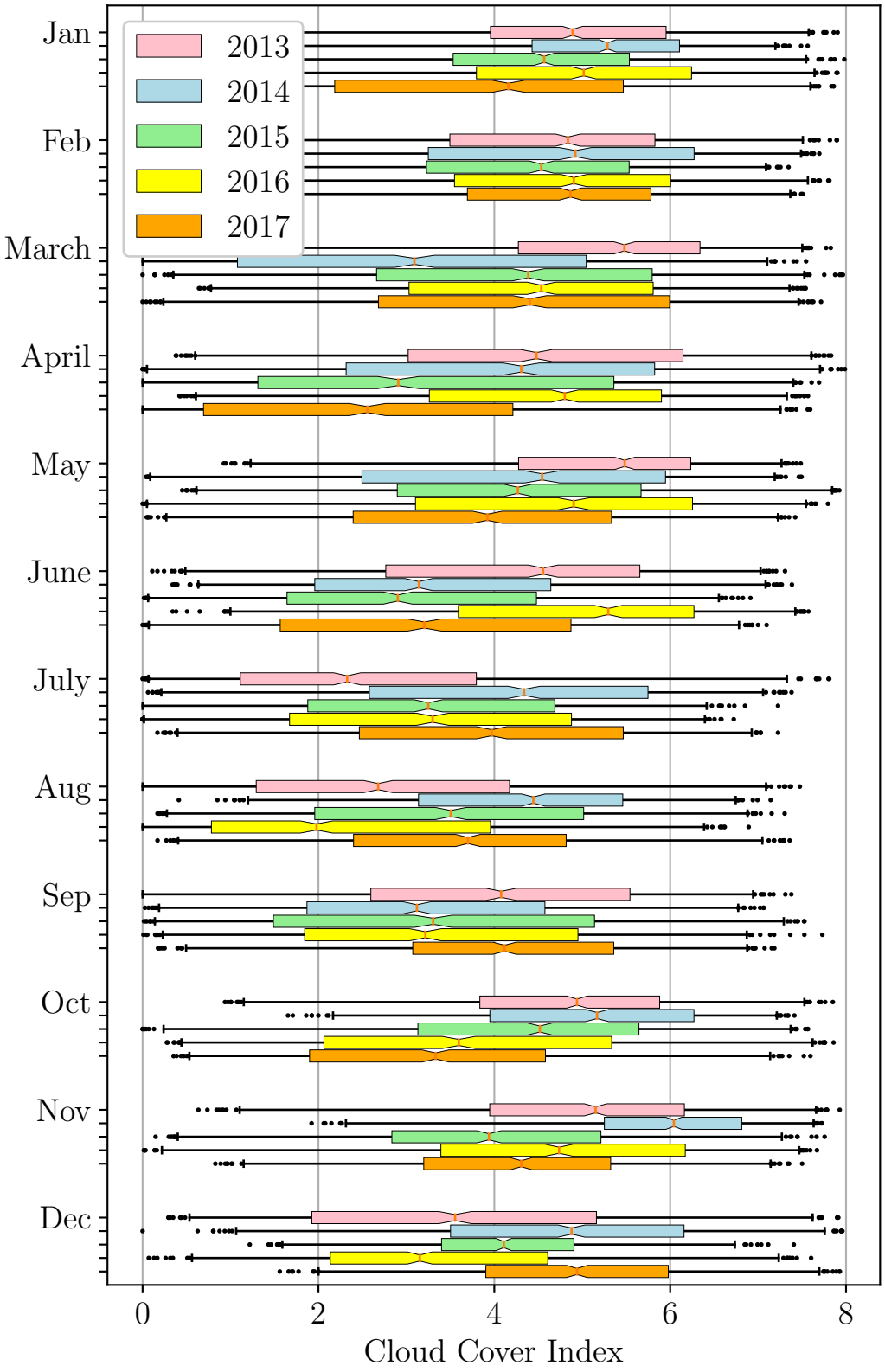


FIGURE F.1: Box plots of the cloud cover index
Box plots of the weighted average cloud cover for each month in the dataset. The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extends from the 1st to the 99th percentiles. Points outside these bounds are plotted individually.

F.1.2 Conditional distributions of the national load

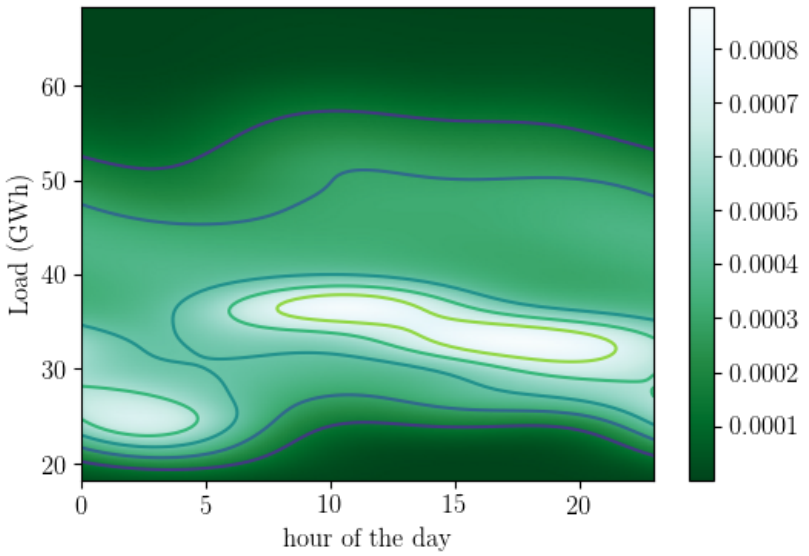


FIGURE F.2: Load conditioned on the hour of the day
Empirical distribution of the national load conditionally on the hour of the day .

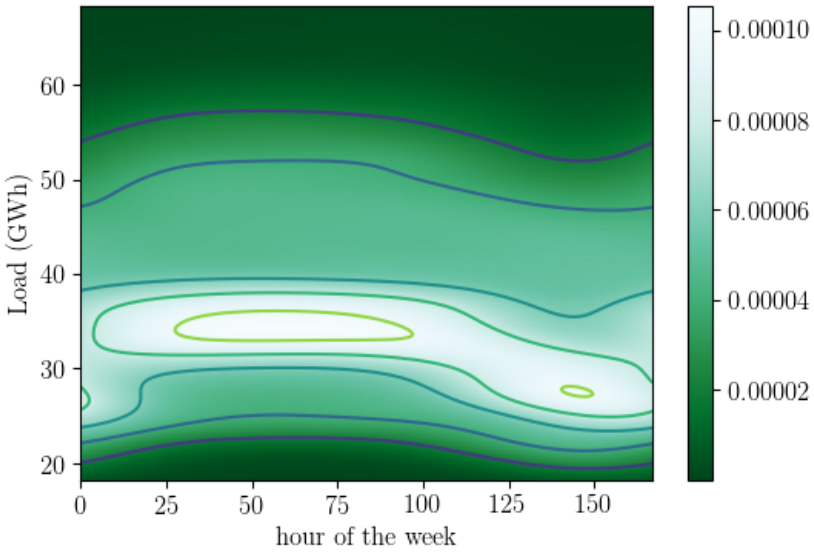


FIGURE F.3: Load conditioned on the hour of the week
Empirical distribution of the national load conditionally on the hour of the week. The variations within a day visible in Figure F.2 cannot be seen here because the window size used with kernel density estimation is too large in this graph.

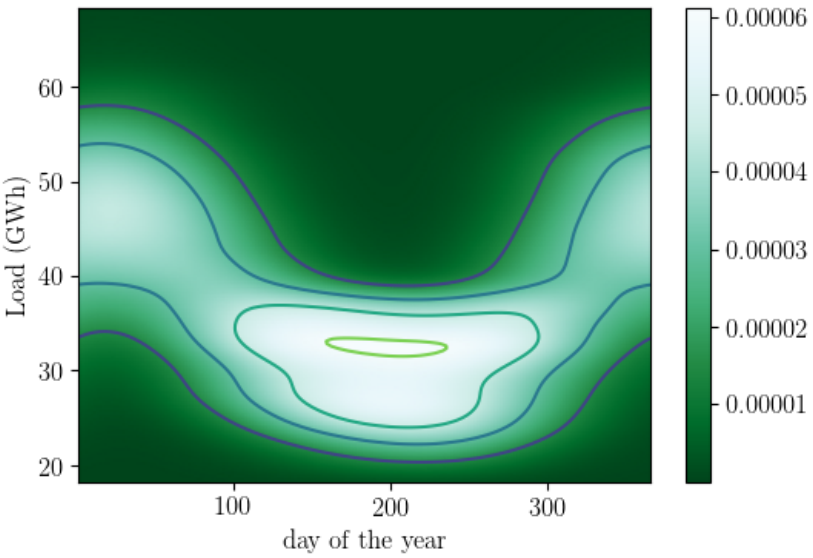


FIGURE F.4: Load conditioned on the day of the year
Empirical distribution of the national load conditionally on the day of the year.

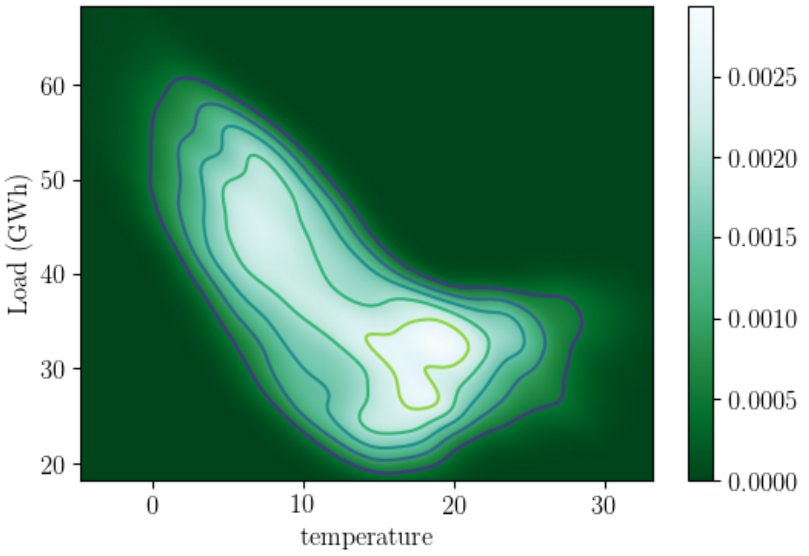


FIGURE F.5: Load conditioned on the averaged temperature
Empirical distribution of the national load conditionally on the value of the average temperature.

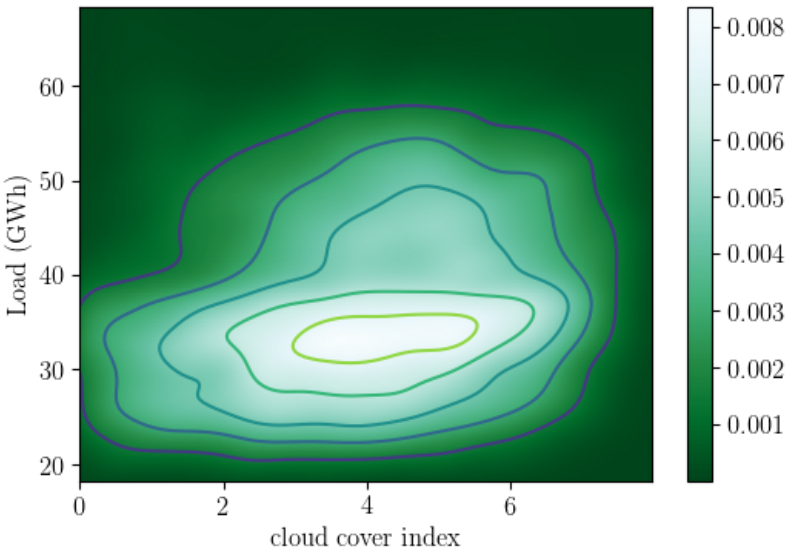


FIGURE F.6: Load conditioned on the cloud cover index
Empirical distribution of the load conditionally on the value of the average cloud cover index.

F.1.3 Notable bivariate conditional expectations

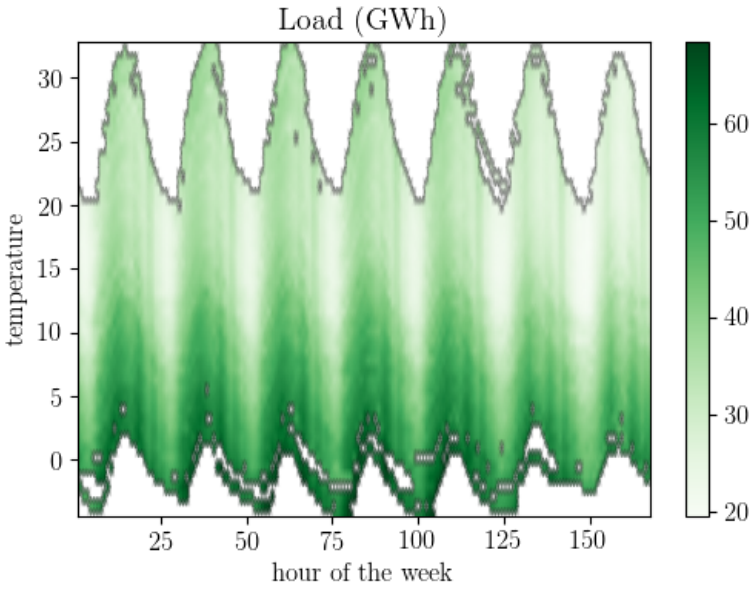


FIGURE F.7: Load conditioned on hours and temperatures
Expectation of the nationally aggregated load conditioned on the hour of the week and the average temperature. Blank parts correspond to low-density regions of the input space.

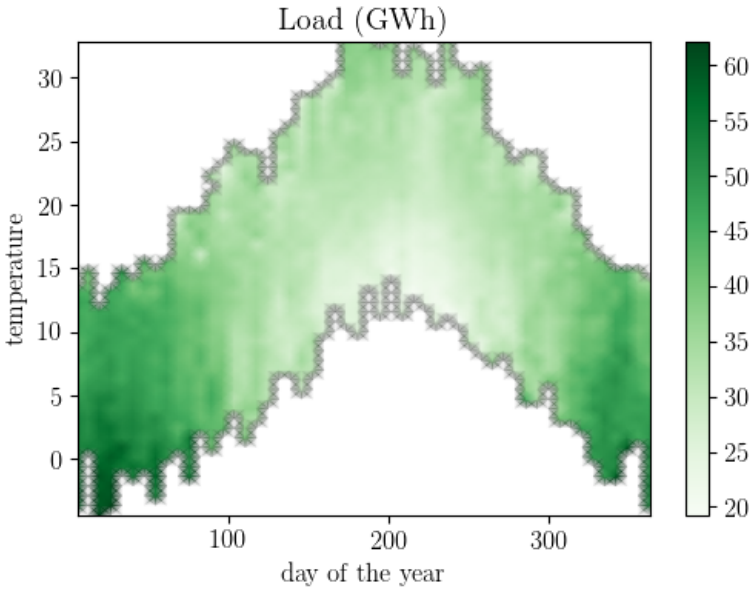


FIGURE F.8: Load conditioned on temperatures and days
Expectation of the nationally aggregated load conditioned on the average temperature and the day of the year.

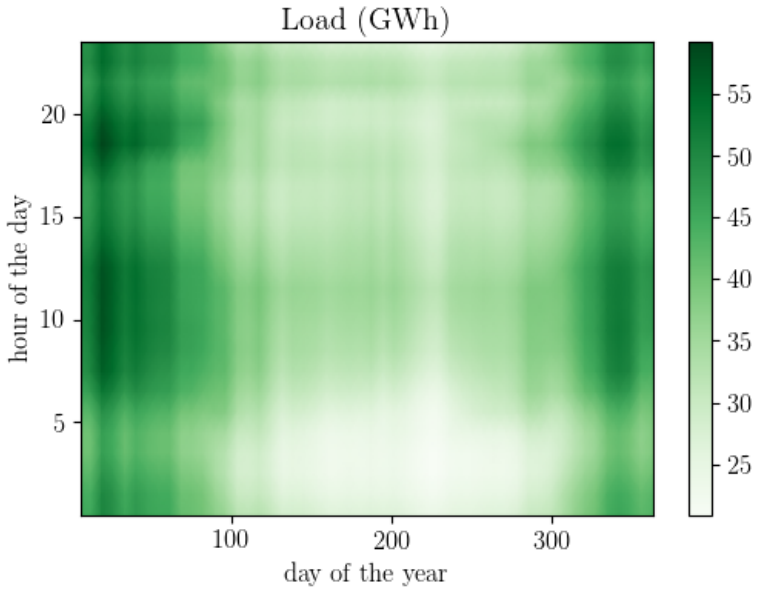


FIGURE F.9: Load conditioned on hours and days of the year
Expectation of the nationally aggregated load conditioned on the hour of the day and the day of the year.

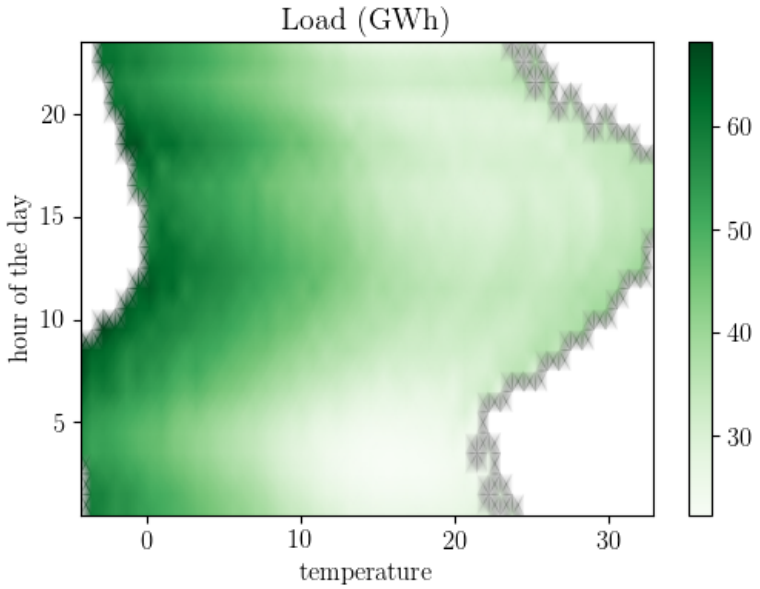


FIGURE F.10: Load conditioned on hours and temperatures
Expectation of the nationally aggregated load conditioned on the hour of the day and the average temperature.

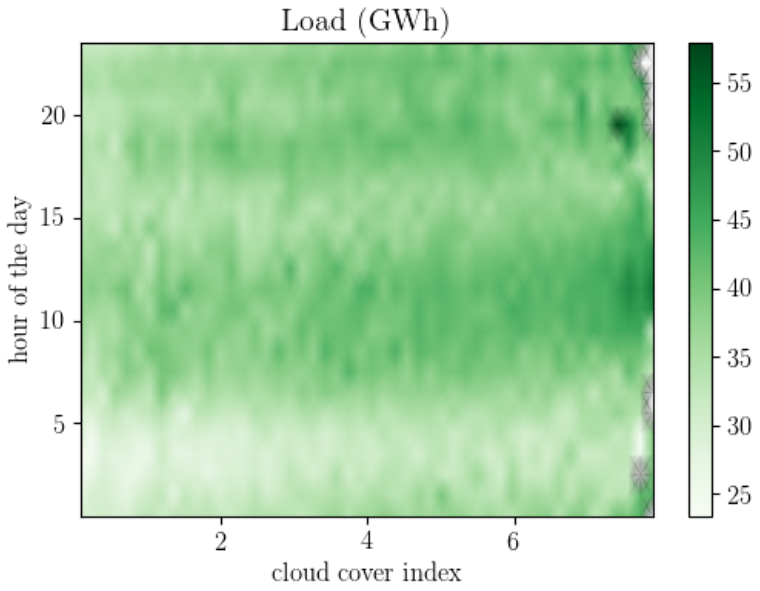


FIGURE F.11: Load conditioned on the hour and the cloud cover
Expectation of the nationally aggregated load conditioned on the hour of the day and the average cloud cover index.

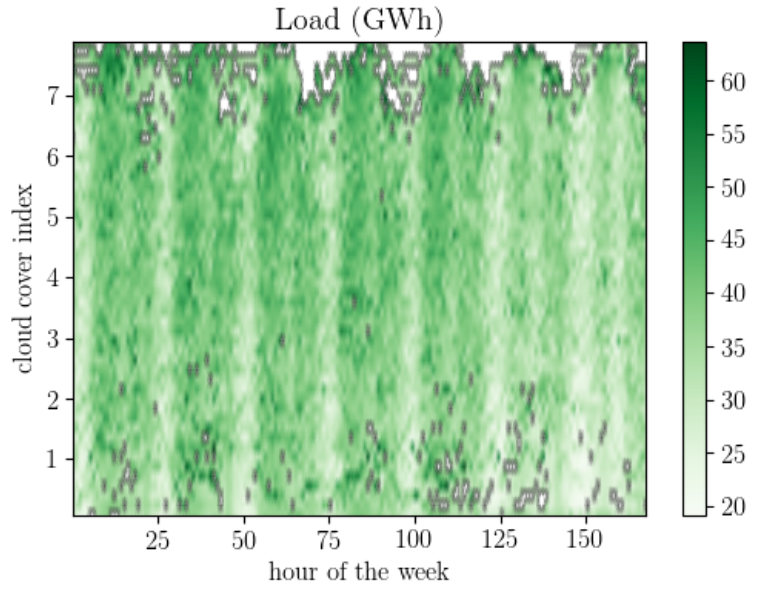


FIGURE F.12: Load conditioned on hours and cloud covers
Expectation of the nationally aggregated load conditioned on the hour of the week and the average cloud cover index. The blank parts correspond to low density regions.

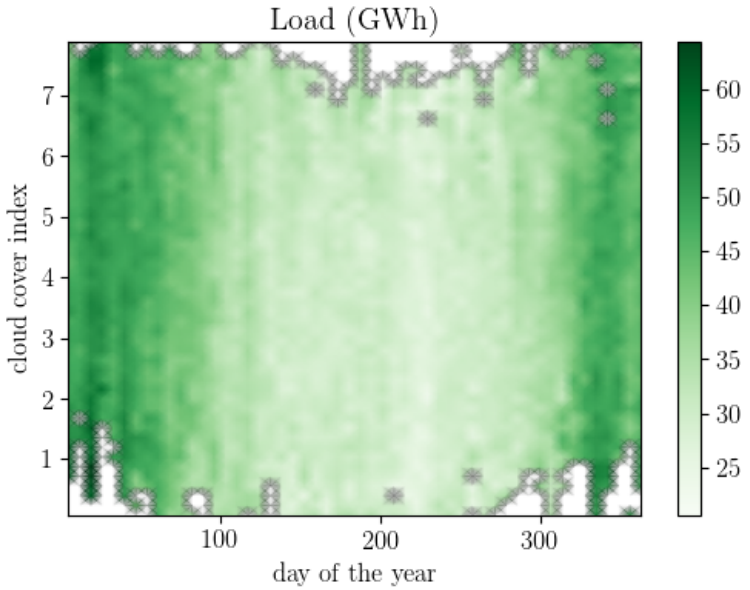


FIGURE F.13: Load conditioned on days and the cloud cover
Expectation of the nationally aggregated load conditioned on the average cloud cover index and the day of the year.

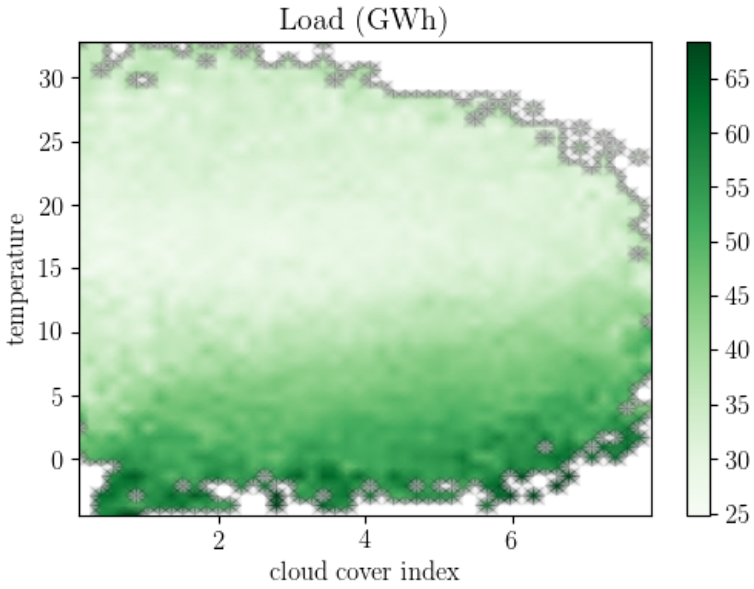


FIGURE F.14: Load conditioned on weather conditions
Expectation of the nationally aggregated load conditioned on the average temperature and the average cloud cover index.

F.1.4 Erratic local loads

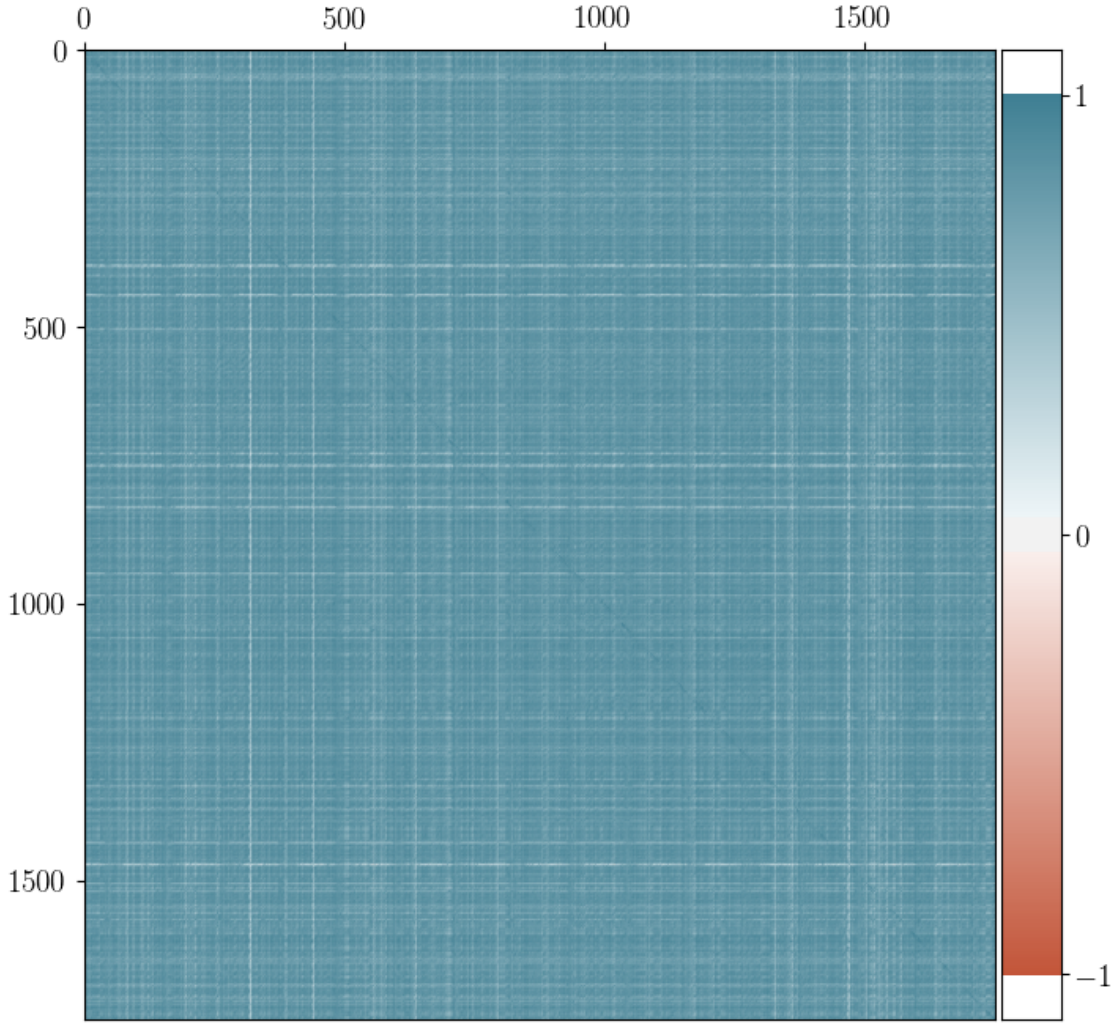


FIGURE F.15: Correlations between the substations
Pairwise correlations of the load of the different substations, computed with the 5-year-long dataset.

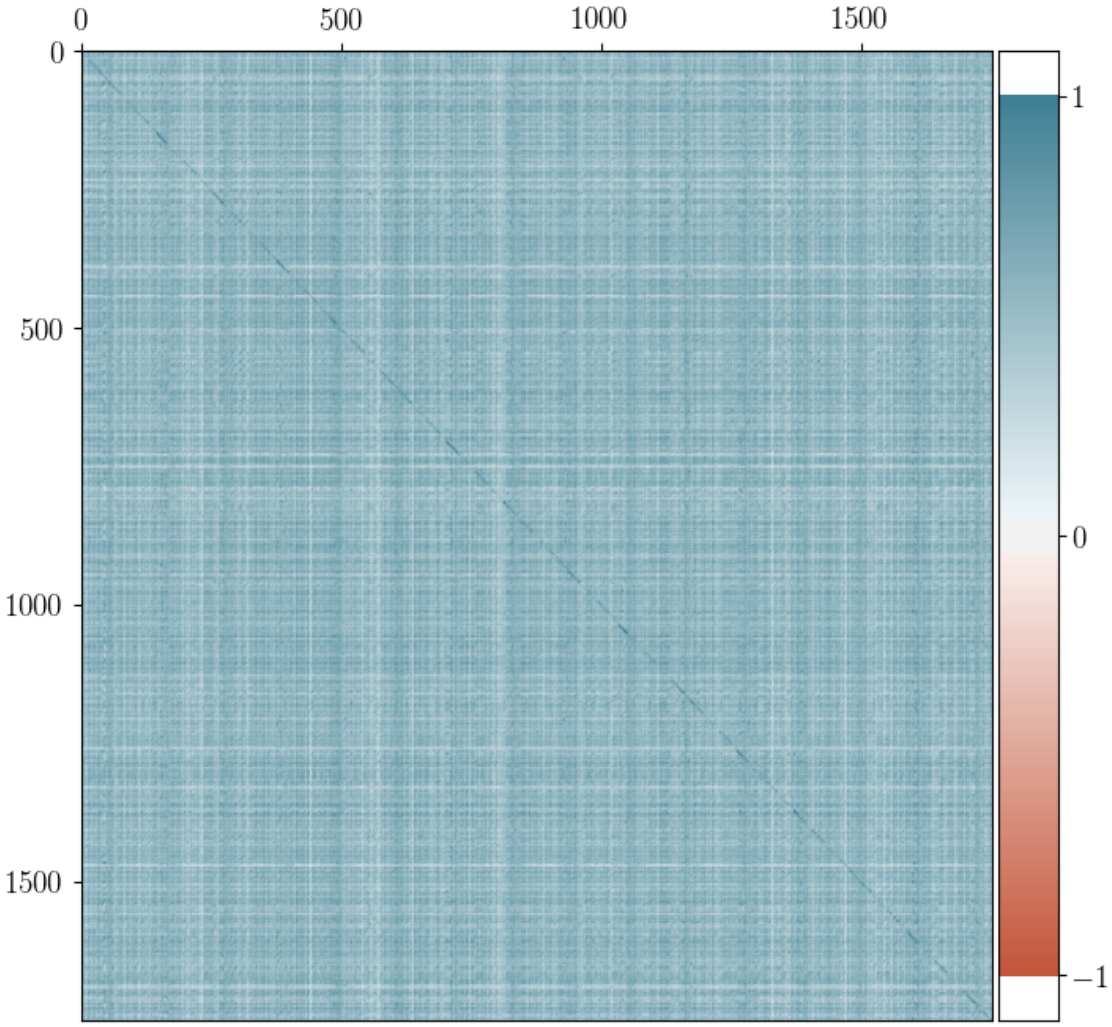


FIGURE F.16: Correlation between the detrended substations
Correlation of the loads at the different substations after an estimate of the load for each day of the year and hour of the week is subtracted. The estimate is obtained with a kernel regression on the observations.

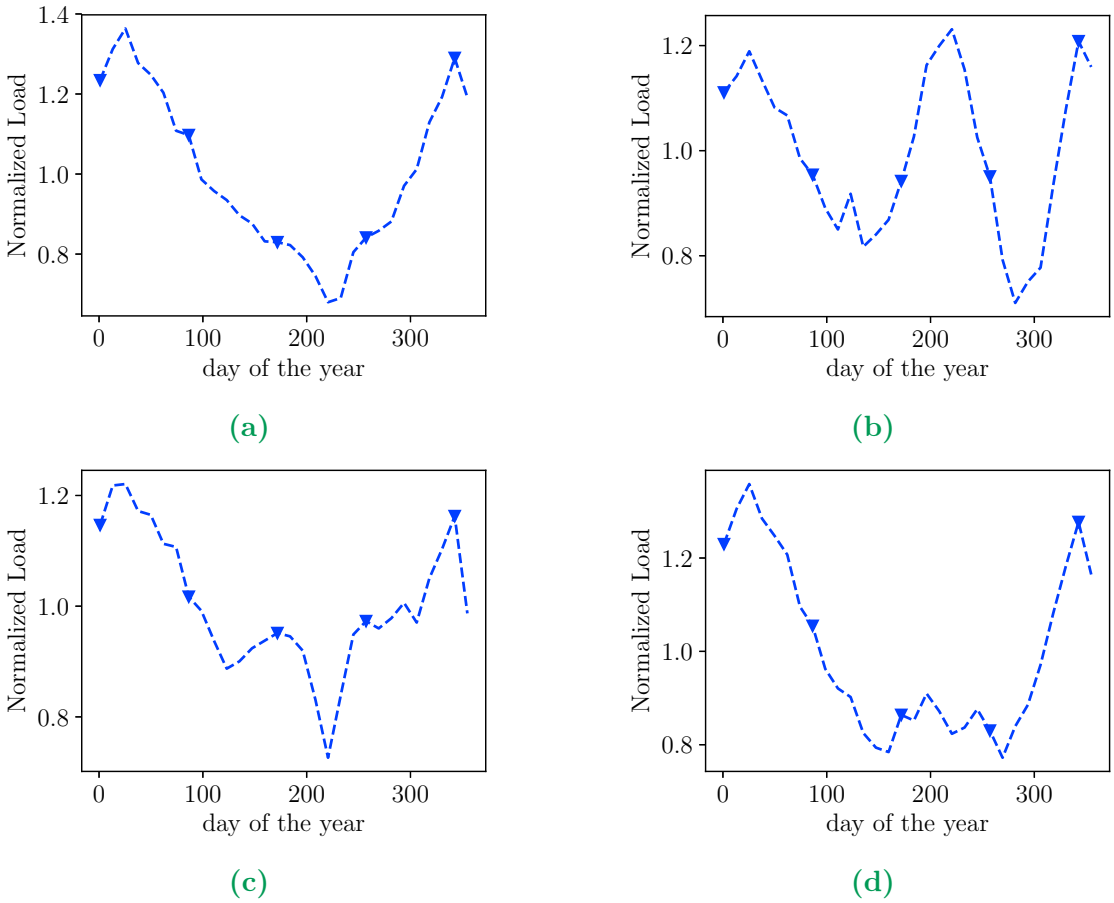


FIGURE F.17: Local loads conditioned on the day of the year
Empirical expectations of the normalized load at 4 different substations conditioned on the day of the year. The load in Figure F.17a is rather similar to the national load presented in Figure 2.11, the 3 others are quite different. There is a relatively larger reduction of the load during the summer in Figure F.17c. On the contrary, Figure F.17b and Figure F.17d that correspond to substations in the Alps and near the French Riviera present a substantial increase of the load during the vacation periods.

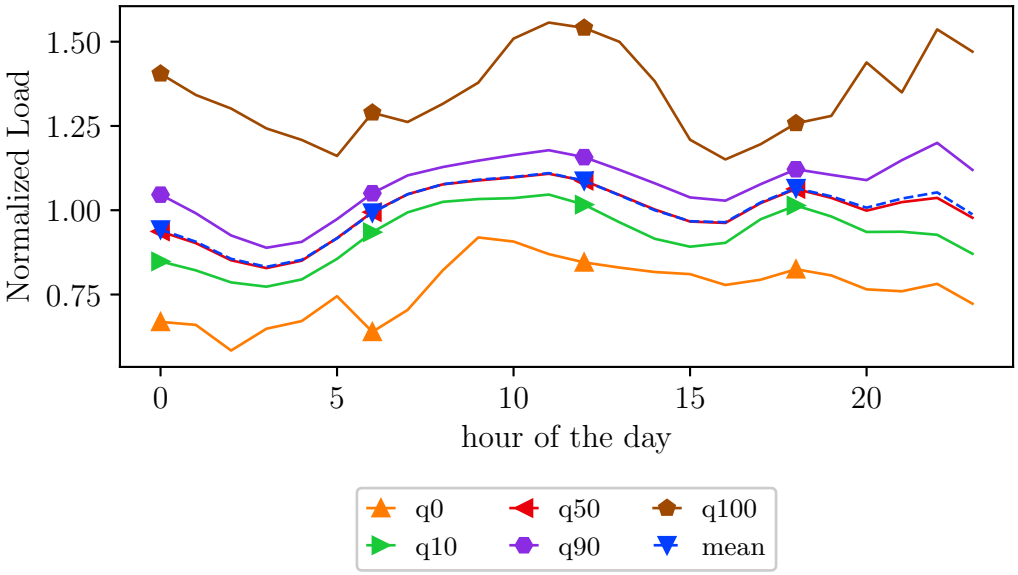


FIGURE F.18: Local loads conditioned on the hour of the day
Quantiles over the substations of the expectations conditioned on the hour of the day of the centered and normalized loads.

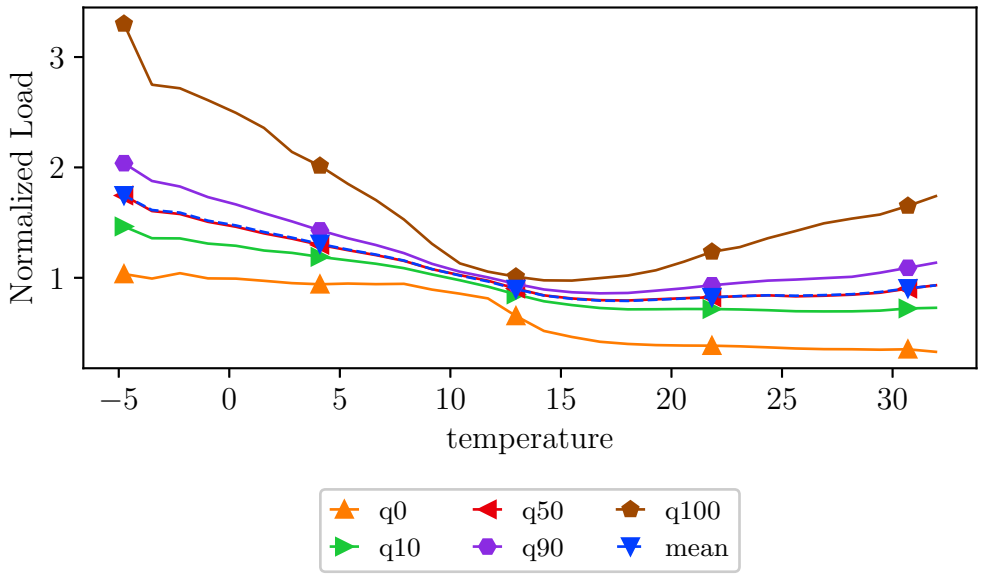


FIGURE F.19: Local loads conditioned on the temperature
Quantiles over the substations of the expectations conditioned on the temperature of the centered normalized loads.

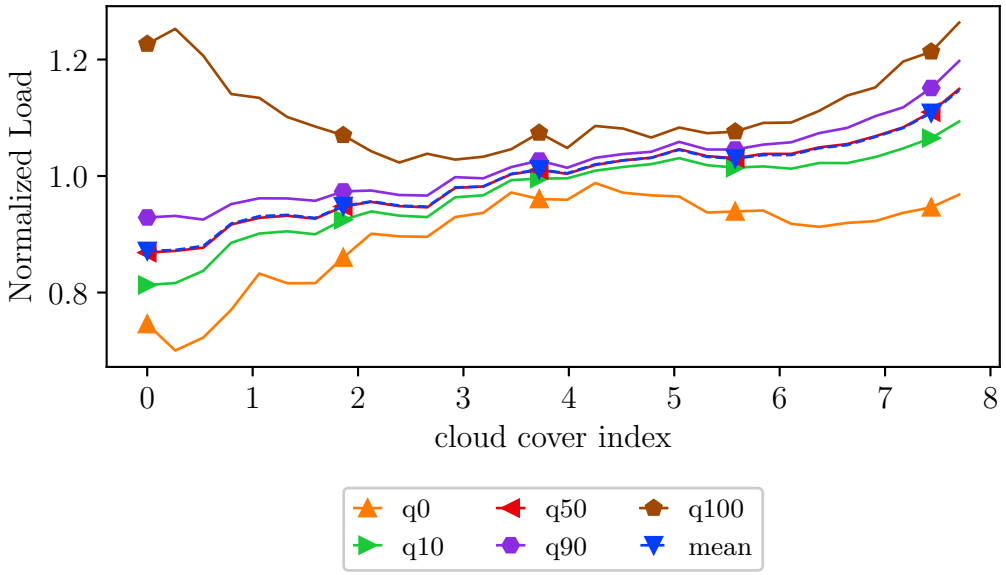


FIGURE F.20: Local loads conditioned on the cloud cover
Quantiles over the substations of the expectations conditioned on the cloud covers of the centered normalized loads.

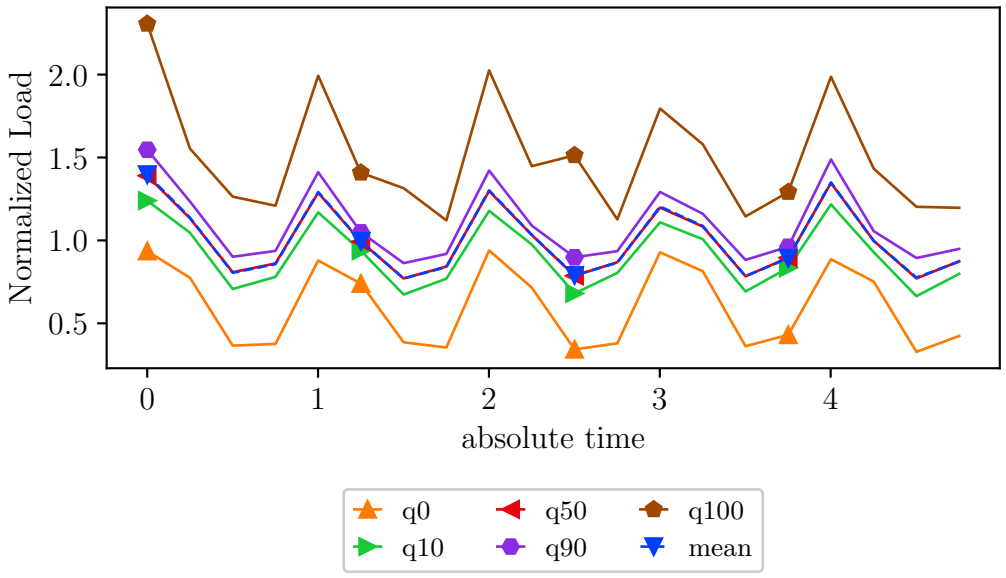


FIGURE F.21: Quantiles of the smoothed local loads smoothed
Quantiles over the substations of the centered normalized loads smoothed over the 5 years in the database.

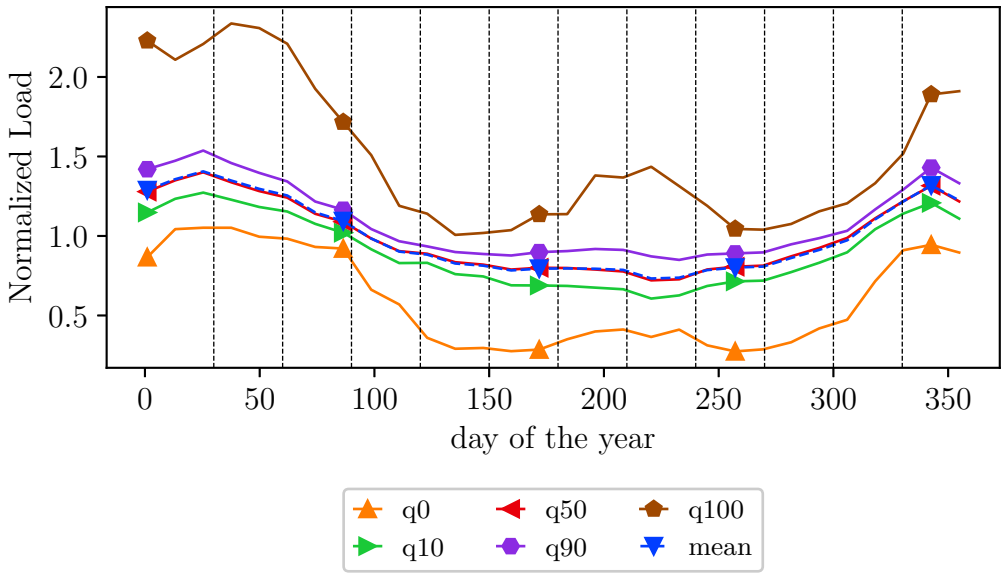


FIGURE F.22: Local loads conditioned on the day of the year
Quantiles over the substations of the expectations conditioned on the day of the year of the centered normalized loads. The vertical lines separate the different months.

F.2 Additional Figures and Tables for Chapter 3

F.2.1 Parametrization for the substations

	Category	Name	Symbol
Date and time	cyclic	hour of the week	h
		day of the year	d
	indicators	holidays	1_{hld}
		days before a holiday	1_{hld^-}
		days after a holiday	1_{hld^+}
		Christmas period	1_{xmas}
		Sun is up	1_{sun}
	absolute time	timestamp	t
Weather	acyclic	temperatures	T^s
		δ hours-delayed temperatures	$\mathsf{T}^s_{-\delta}$
		maximum over a δ hours window	$\overline{\mathsf{T}}^s_{-\delta}$
		minimum over a δ hours window	$\underline{\mathsf{T}}^s_{-\delta}$
		cloud covers	c^s
Past loads	acyclic	δ hours-delayed load	$\ell_{-\delta}$

TABLE F.2: Inputs to the short-term load forecasting models
There are 2 copies of the weather-related inputs corresponding to the 2 closest weather station and one copy of the past information for each $\delta \in \{24, 48\}$.

Name	Symbol	Parametrization
hour of the week	h	168 knots
day of the year	d	32 knots
Christmas period	1_{xmas}	indicator
timestamp	t	linear function
temperatures	T^s	9 knots
δh delayed temperatures	$\mathsf{T}^s_{-\delta}$	9 knots
last δh maxima	$\overline{\mathsf{T}}^s_{-\delta}$	9 knots
last δh minima	$\underline{\mathsf{T}}^s_{-\delta}$	9 knots
δh-delayed load	$\ell_{-\delta}$	3 knots

TABLE F.3: Univariate features for the local forecasts
Set $\mathcal{U}$ of univariate features with the corresponding parametrization for the short-term local load forecasting models.

Names	Symbols	Parametrization
cloud covers and day/night	$(\mathsf{c}^s, 1_{\mathsf{sun}})$	3 knots & indicator
hour of the week and holiday	$(\mathsf{h}, 1_{\mathsf{hld}})$	168 knots & indicator
hour of the week and day before a holiday	$(\mathsf{h}, 1_{\mathsf{hld}^-})$	168 knots & indicator
hour of the week and day after a holiday	$(\mathsf{h}, 1_{\mathsf{hld}^+})$	168 knots & indicator
hour of the week and δh-delayed load	$(\mathsf{h}, \ell_{-\delta})$	42 knots & linear
hour of the week and day of the year	(h, d)	168 & 16 knots
temperatures and day of the year	$(\mathsf{T}^s, \mathsf{d})$	9 & 16 knots

TABLE F.4: Bivariate features for the local forecasting models
Set $\mathcal{B}$ of bivariate features with the corresponding parametrization for the short-term load forecasting models.

F.2.2 Estimated univariate effects for the national model

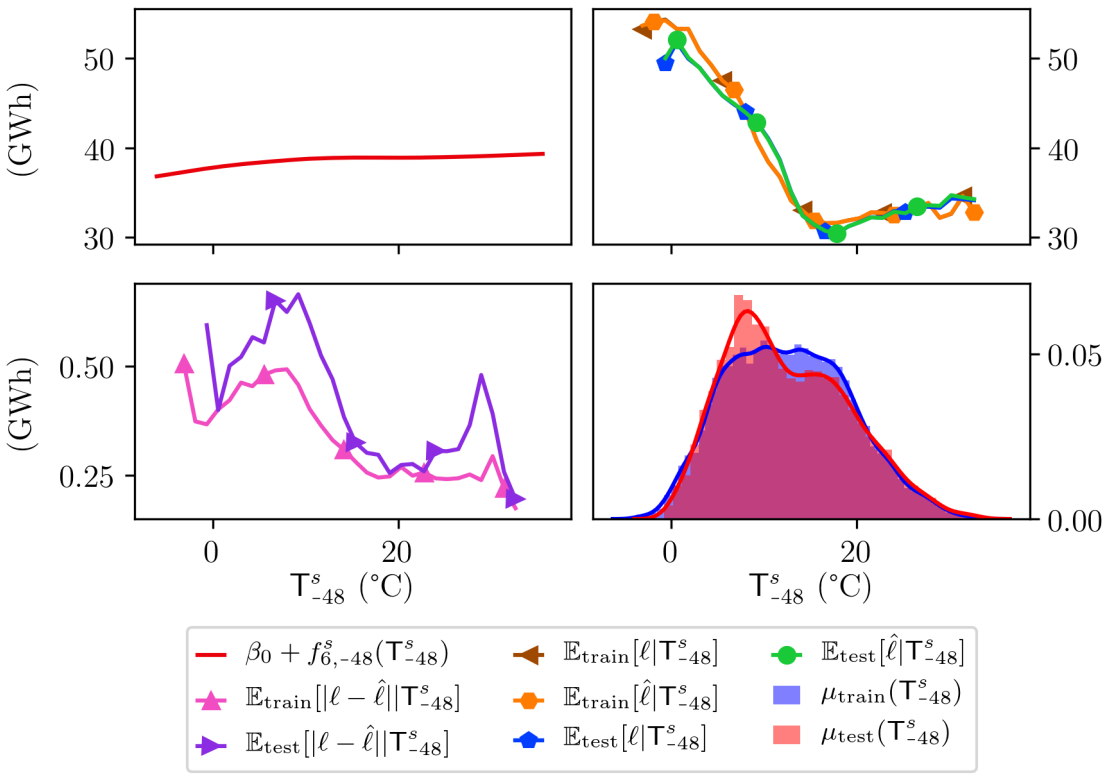


FIGURE F.23: Estimated effect of the 48 h-delayed temperature

(*top left*) Effect $\beta_0 + f^s_{6,\text{-}48}(T^s_{-48})$ learned by the national short-term model for the 48 hours-delayed weighted temperature defined with Table F.1.

(*bottom left*) Norm of the conditional residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||T^s_{-48}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||T^s_{-48}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|T^s_{-48}]$ and $\mathbb{E}_{\text{test}}[\ell|T^s_{-48}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|T^s_{-48}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|T^s_{-48}]$.

(*bottom left*) Density of the data in the training and the test sets.

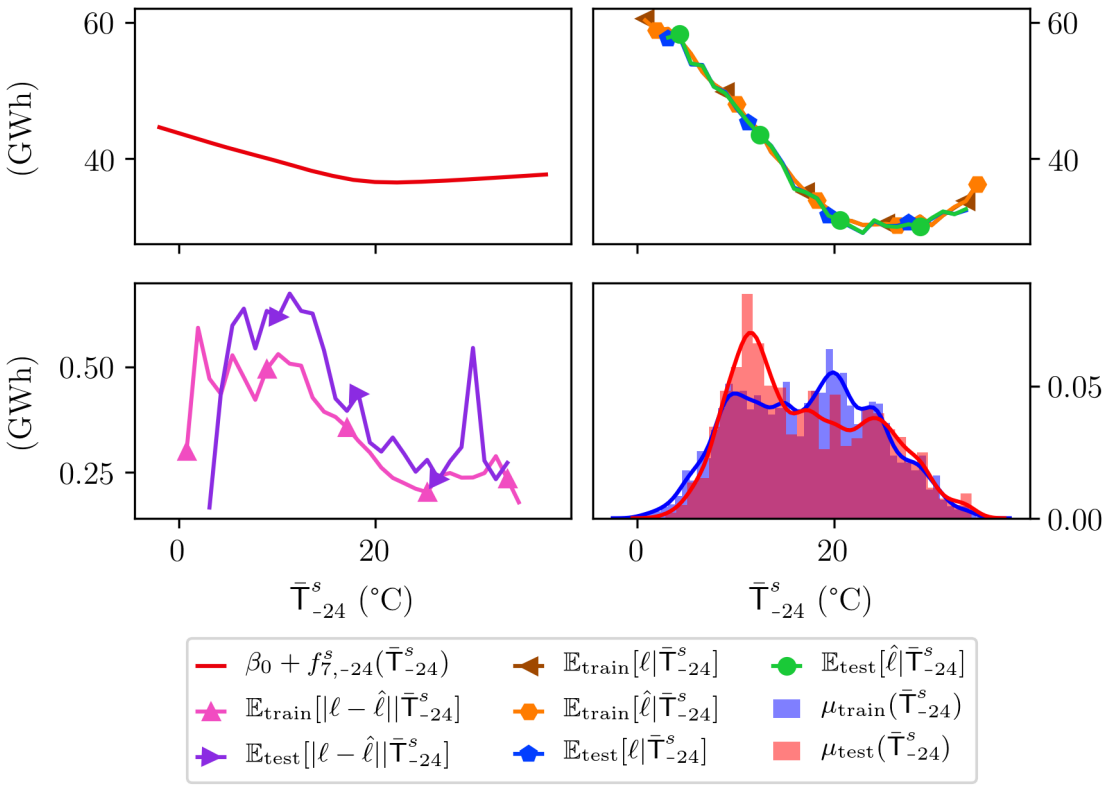


FIGURE F.24: Estimated effect of the 24 h max temperature

(*top left*) Effect $\beta_0 + f^s_{7,\text{-}\delta}(\bar{\mathsf{T}}^s_{\text{-}24})$ learned by the national short-term model for the maximum over 24 hours of the weighted temperature defined with Table F.1.

(*bottom left*) Norm of the conditional residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\bar{\mathsf{T}}^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\bar{\mathsf{T}}^s_{\text{-}24}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\bar{\mathsf{T}}^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[\ell|\bar{\mathsf{T}}^s_{\text{-}24}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\bar{\mathsf{T}}^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\bar{\mathsf{T}}^s_{\text{-}24}]$.

(*bottom left*) Density of the data in the training and the test sets.

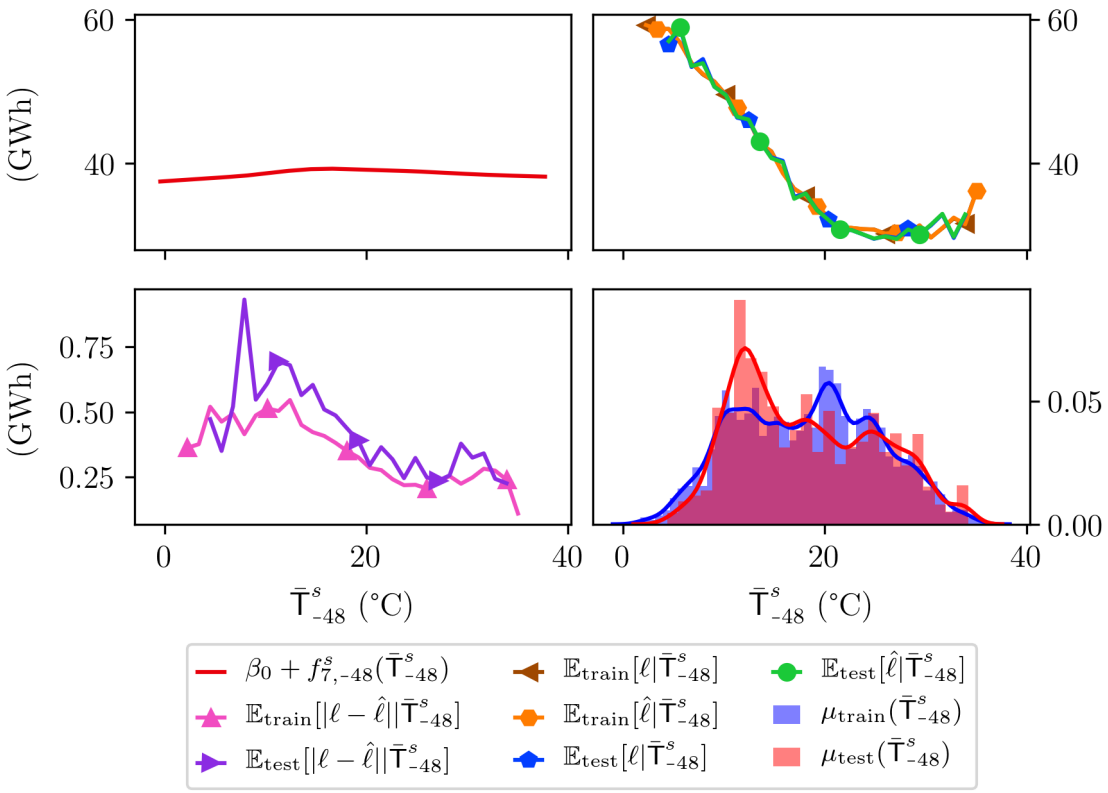


FIGURE F.25: Estimated effect of the 48 h max temperature

(*top left*) Effect $\beta_0 + f^s_{7,\text{-}\delta}(\bar{\mathsf{T}}^s_{\text{-}48})$ learned by the national short-term model for the maximum over 48 hours of the weighted temperature defined with Table F.1.

(*bottom left*) Norm of the conditional residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\bar{\mathsf{T}}^s_{\text{-}48}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\bar{\mathsf{T}}^s_{\text{-}48}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\bar{\mathsf{T}}^s_{\text{-}48}]$ and $\mathbb{E}_{\text{test}}[\ell|\bar{\mathsf{T}}^s_{\text{-}48}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\bar{\mathsf{T}}^s_{\text{-}48}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\bar{\mathsf{T}}^s_{\text{-}48}]$.

(*bottom left*) Density of the data in the training and the test sets.

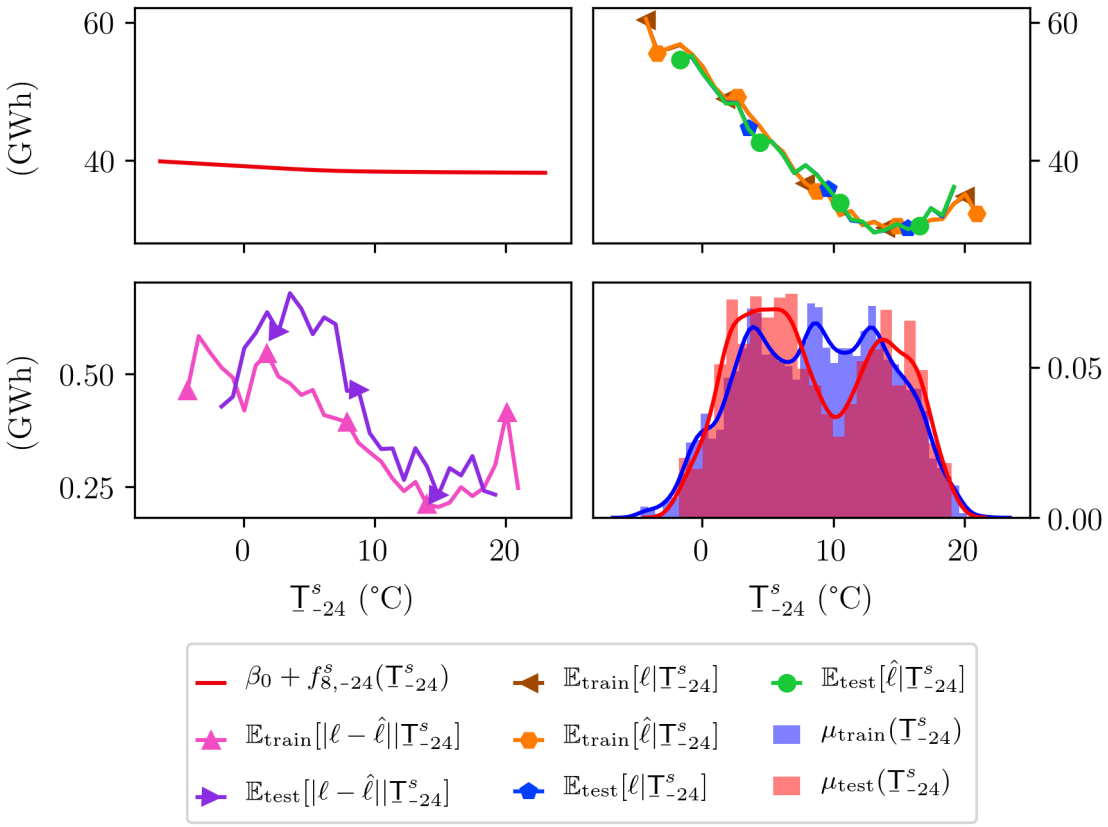


FIGURE F.26: Estimated effect of the 24 h min temperature

(*top left*) Effect $\beta_0 + f^s_{8,\text{-}24}(\underline{\mathrm{T}}^s_{\text{-}24})$ learned by the national short-term model for the minimum over 24 hours of the weighted temperature defined with Table F.1.

(*bottom left*) Norm of the conditional residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\underline{\mathrm{T}}^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\underline{\mathrm{T}}^s_{\text{-}24}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\underline{\mathrm{T}}^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[\ell|\underline{\mathrm{T}}^s_{\text{-}24}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\underline{\mathrm{T}}^s_{\text{-}24}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\underline{\mathrm{T}}^s_{\text{-}24}]$.

(*bottom left*) Density of the data in the training and the test sets.

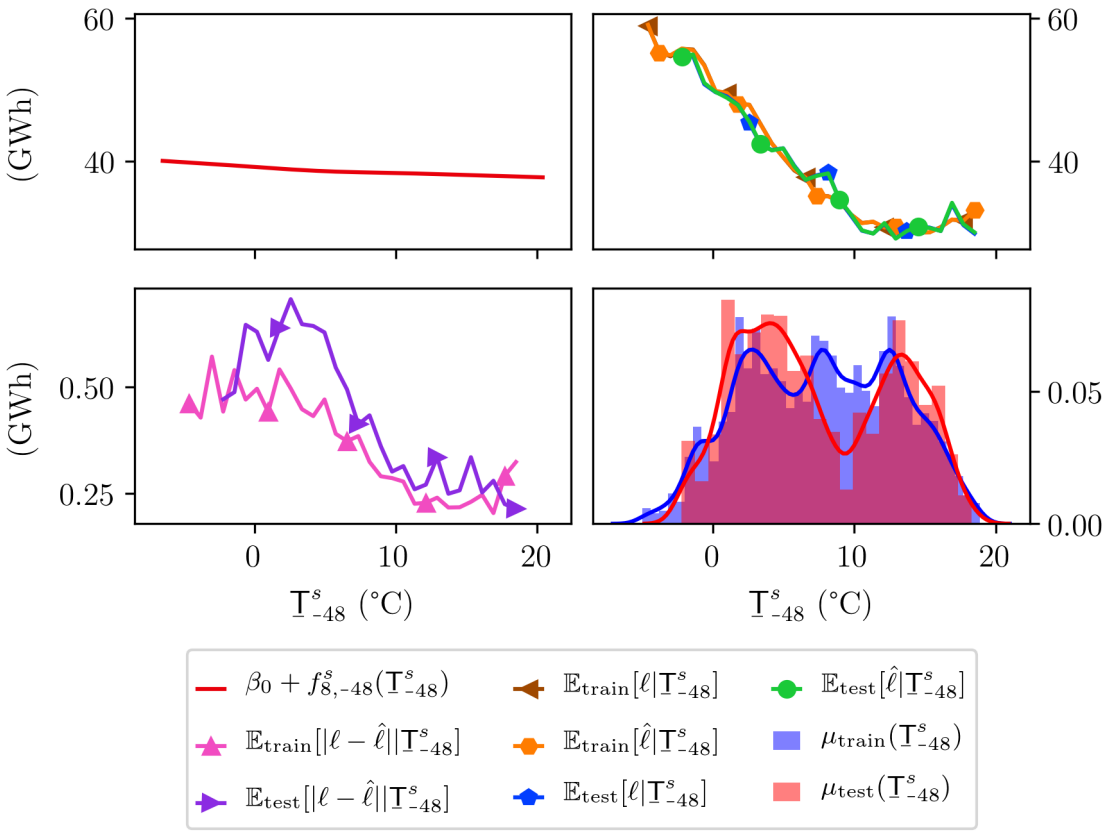


FIGURE F.27: Estimated effect of the 48 h min temperature

(*top left*) Effect $\beta_0 + f^s_{8,\text{-}48}(\underline{\mathrm{T}}^s_{\text{-}48})$ learned by the national short-term model for the minimum over 48 hours of the weighted temperature defined with Table F.1.

(*bottom left*) Norm of the conditional residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\underline{\mathrm{T}}^s_{\text{-}48}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\underline{\mathrm{T}}^s_{\text{-}48}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\underline{\mathrm{T}}^s_{\text{-}48}]$ and $\mathbb{E}_{\text{test}}[\ell|\underline{\mathrm{T}}^s_{\text{-}48}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\underline{\mathrm{T}}^s_{\text{-}48}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\underline{\mathrm{T}}^s_{\text{-}48}]$.

(*bottom left*) Density of the data in the training and the test sets.

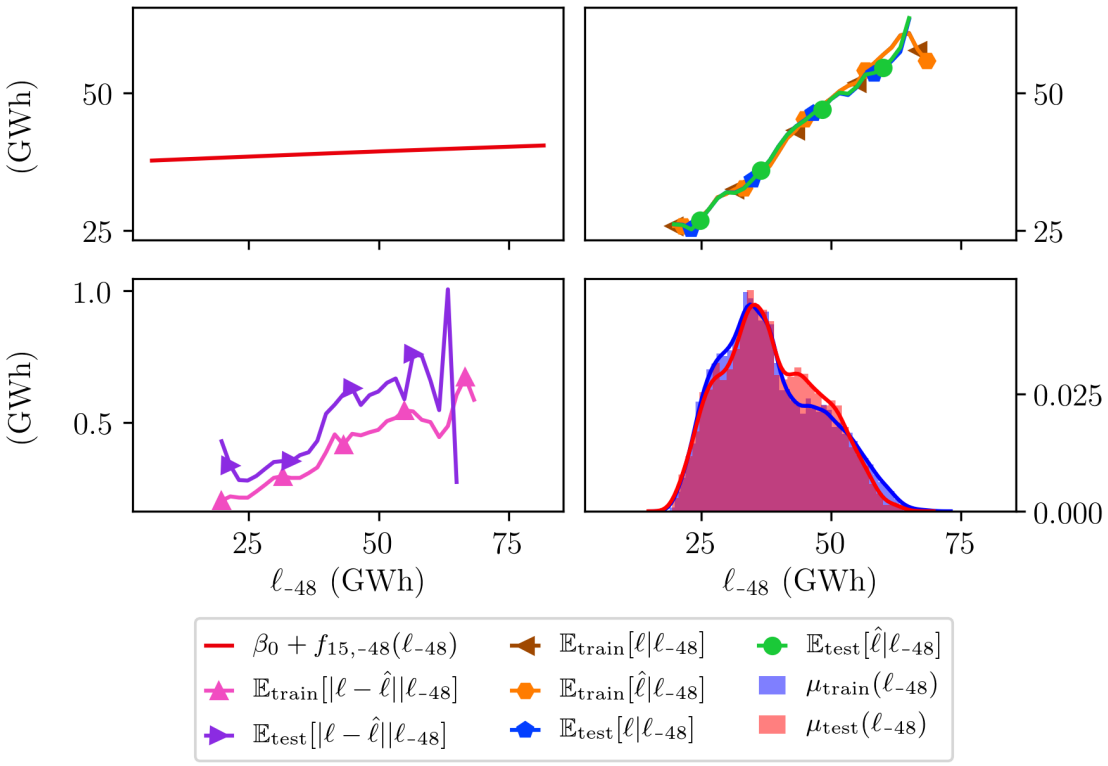


FIGURE F.28: Estimated effect of the 48 h-delayed load

(*top left*) Effect $\beta_0 + f_{15,\text{-}48}(\ell_{-48})$ learned by the national short-term model for the 48 hours-delayed load.

(*bottom left*) Norm of the conditional residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||\ell_{-48}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||\ell_{-48}]$.

(*top right*) Conditional loads $\mathbb{E}_{\text{train}}[\ell|\ell_{-48}]$ and $\mathbb{E}_{\text{test}}[\ell|\ell_{-48}]$ with the conditional forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|\ell_{-48}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|\ell_{-48}]$.

(*bottom left*) Density of the data in the training and the test sets.

F.2.3 Estimated bivariate effects for the national model

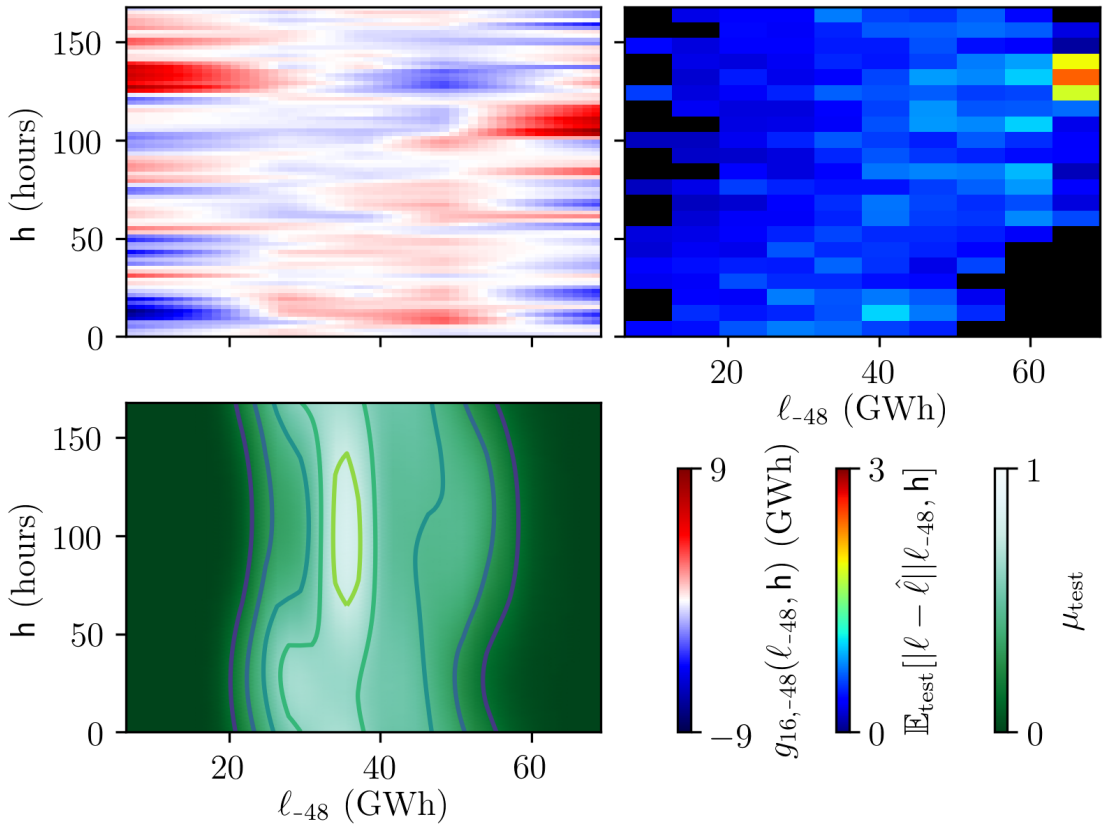


FIGURE F.29: Interaction between past loads and week hours

(*top left*) Interaction $g_{16,\text{-}48}(\mathsf{h}, \ell_{\text{-}48})$ of the past load and the hour of the week in the short-term national model.

(*bottom left*) Estimated marginal density $\mu_{\text{test}}(\mathsf{h}, \ell_{\text{-}48})$ of the inputs in the test set.

(*top right*) Average residuals in the test set $\mathbb{E}_{\text{test}}[\,|\ell - \hat{\ell}|\,|\mathsf{h}, \ell_{\text{-}48}]$.

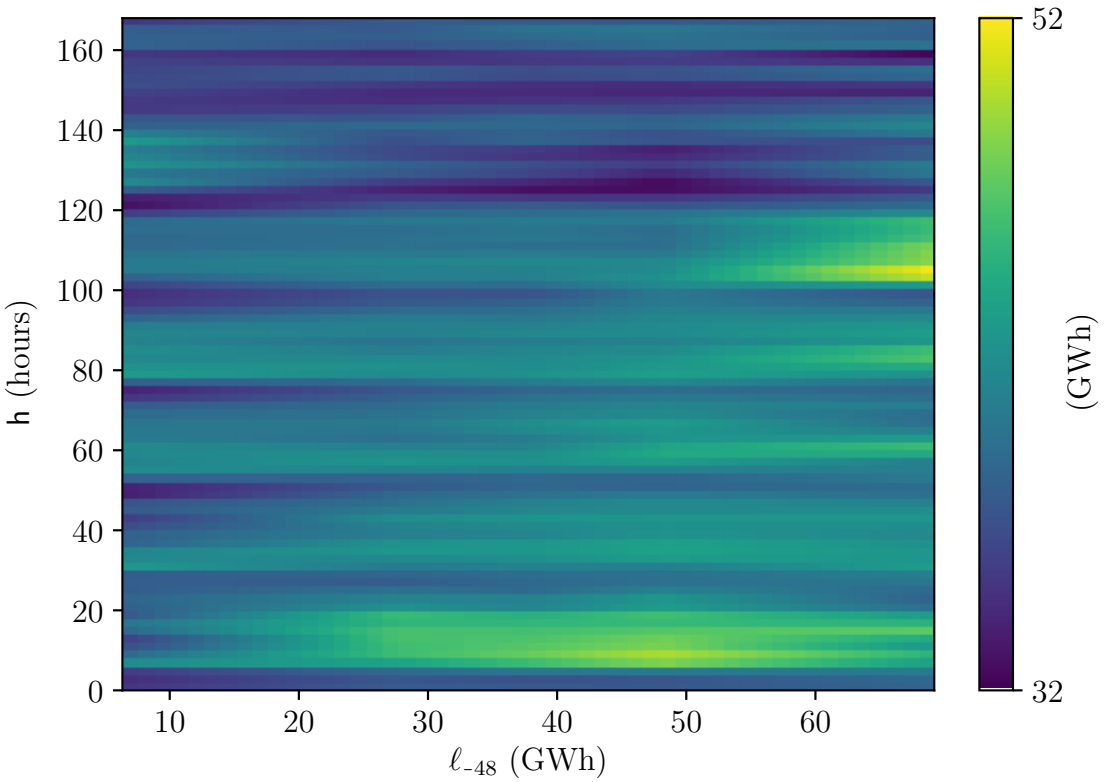


FIGURE F.30: Total effect of the 48 h-delayed load
Sum of the effect related to the 48 hours-delayed loads and the hour of the week $\beta_0 + f_3(\mathsf{h}) + f_{15,\text{-}48}(\ell_{\text{-}48}) + g_{16,\text{-}48}(\mathsf{h}, \ell_{\text{-}48})$ in the short-term national model.

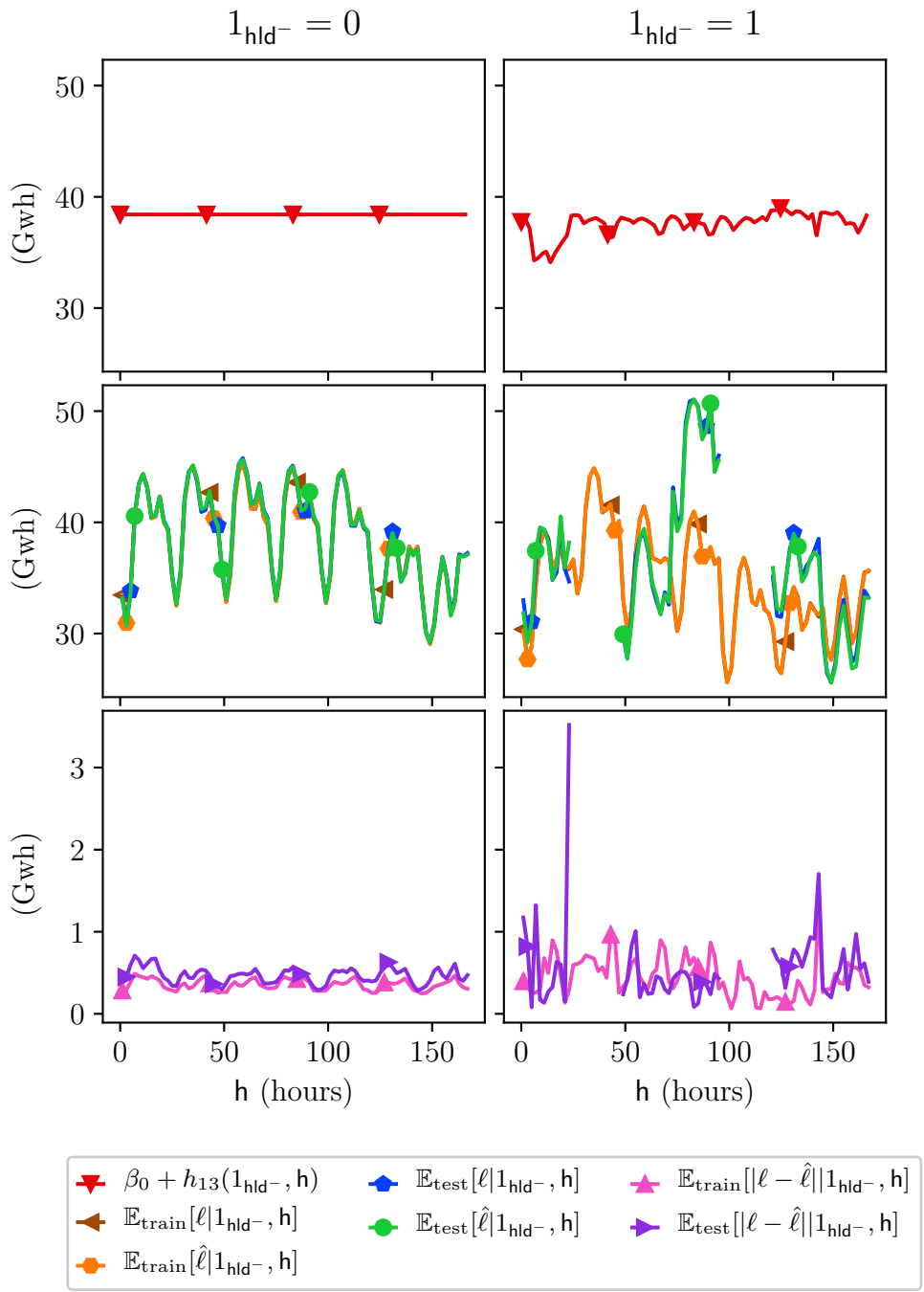


FIGURE F.31: Interaction between hours and coming holidays

(1^{st} *row*) Interaction $\beta_0 + 1_{\text{hld}^-} h_{13}(\mathsf{h})$ between the indicator of the days before a holiday and the hour of the week.

(2^{nd} *row*) Target loads $\mathbb{E}_{\text{train}}[\ell|1_{\text{hld}^-}, \mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\ell|1_{\text{hld}^-}, \mathsf{h}]$ with the forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|1_{\text{hld}^-}, \mathsf{h}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|1_{\text{hld}^-}, \mathsf{h}]$.

(3^{rd} *row*) Marginal norm of the residuals $\mathbb{E}_{\text{train}}[|\ell - \hat{\ell}||1_{\text{hld}^-}, \mathsf{h}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||1_{\text{hld}^-}, \mathsf{h}]$.

The marginal loads, forecasts and residuals are incomplete in the column $1_{\text{hld}^-} = 1$ because there was no holiday on Wednesdays and Saturdays in 2016.

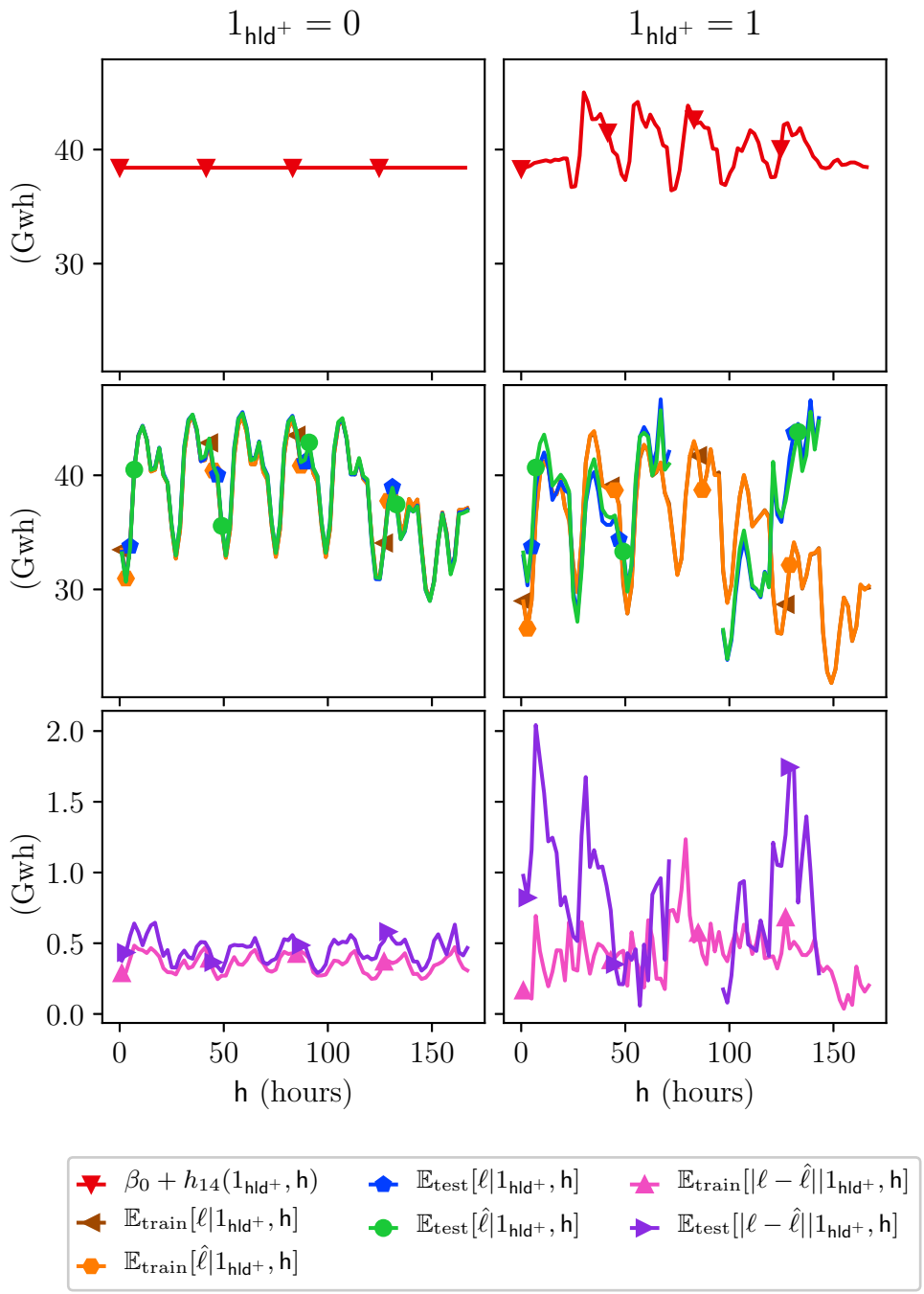


FIGURE F.32: Interaction between week hours and past holidays

(1^{st} *row*) Interaction $\beta_0 + 1_{\text{hld}^+} h_{14}(\text{h})$ between the indicator of the days after a holiday and the hour of the week.

(2^{nd} *row*) Target loads $\mathbb{E}_{\text{train}}[\ell|1_{\text{hld}^+}, \text{h}]$ and $\mathbb{E}_{\text{test}}[\ell|1_{\text{hld}^+}, \text{h}]$ with the forecasts $\mathbb{E}_{\text{train}}[\hat{\ell}|1_{\text{hld}^+}, \text{h}]$ and $\mathbb{E}_{\text{test}}[\hat{\ell}|1_{\text{hld}^+}, \text{h}]$.

(3^{rd} *row*) Marginal norm of the residuals $\mathbb{E}_{\text{train}}[\,|\ell - \hat{\ell}|\,|1_{\text{hld}^+}, \text{h}]$ and $\mathbb{E}_{\text{test}}[|\ell - \hat{\ell}||1_{\text{hld}^+}, \text{h}]$.

The marginal loads, forecasts and residuals are incomplete in the column $1_{\text{hld}^+} = 1$ because there was no holiday on Wednesdays and Saturdays in 2016. Our interpretation for the upper right plot is that a day following a holiday has to compensate for the decrease of the load that impacts the effect of the past load.

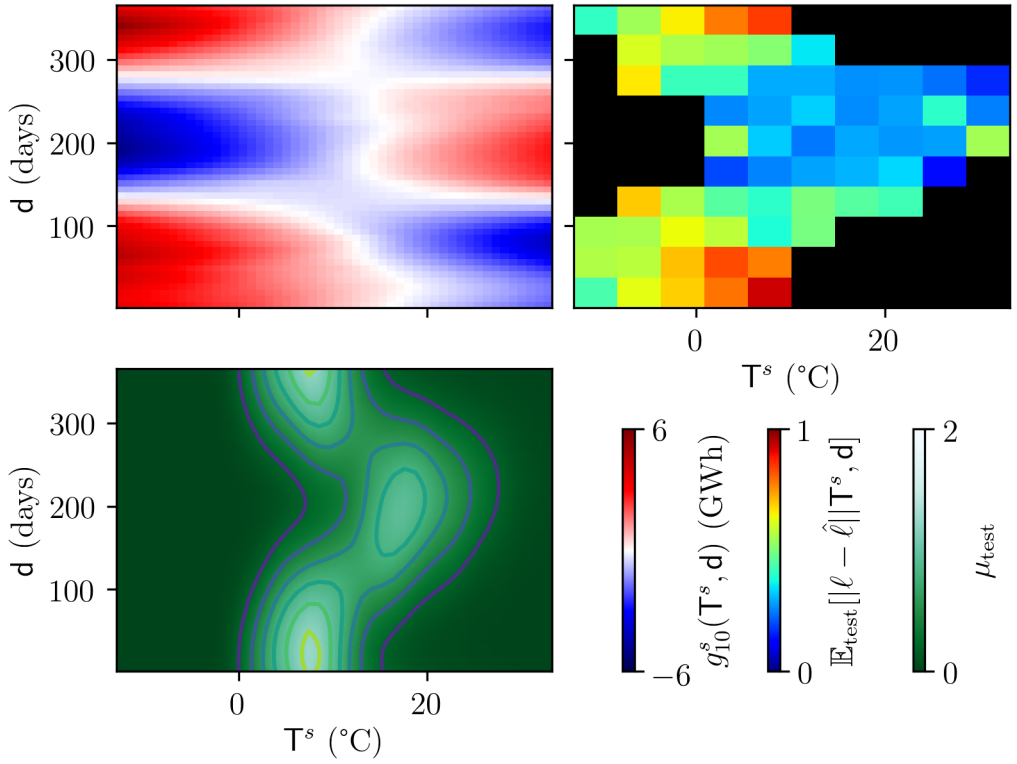


FIGURE F.33: Interaction between year days and temperatures

(*top left*) Interaction $g^s_{10}(\mathsf{T}^s, \mathsf{d})$ between the past load and the hour of the week in the short-term national model.

(*bottom left*) Estimated marginal density $\mu_{\text{test}}(\mathsf{T}^s, \mathsf{d})$ of the inputs in the test set.

(*top right*) Average residuals in the test set $\mathbb{E}_{\text{test}}[\,|\mathsf{y} - \hat{\mathsf{y}}|\,|\mathsf{T}^s, \mathsf{d}]$.

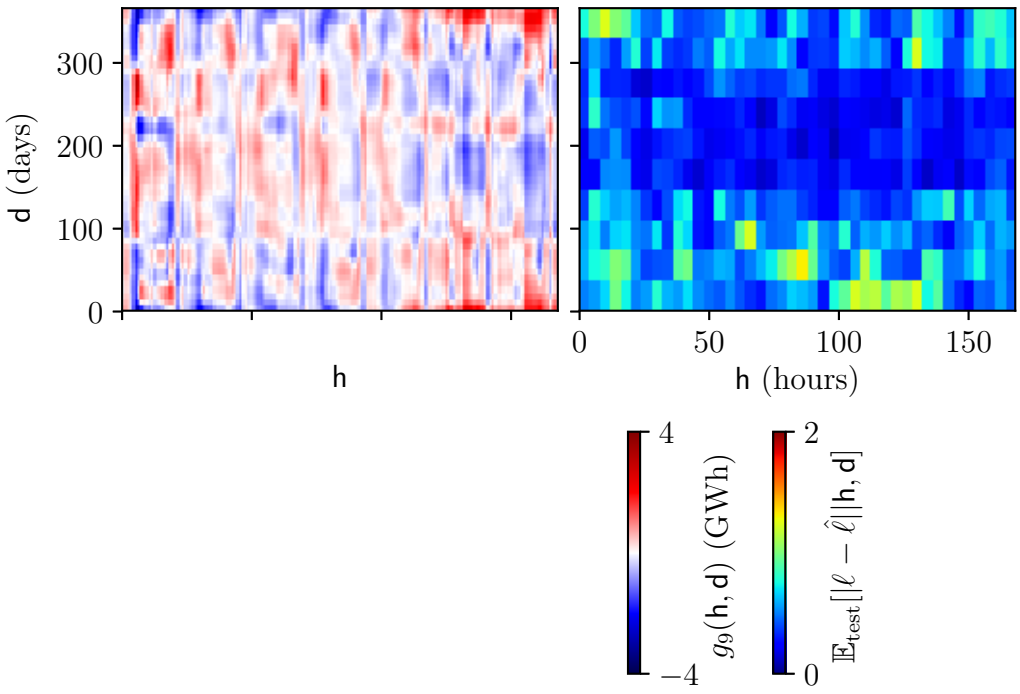


FIGURE F.34: Interaction between hours and days of the year
(*left*) Interaction $g_9(\mathsf{h},\mathsf{d})$ between the hour of the week h and the day of the year d in the short-term national model.
(*right*) Average residuals in the test set $\mathbb{E}_{\text{test}}[\,|\mathsf{y} - \hat{\mathsf{y}}|\,|\mathsf{h},\mathsf{d}]$.

F.2.4 Quantiles of the local univariate effects

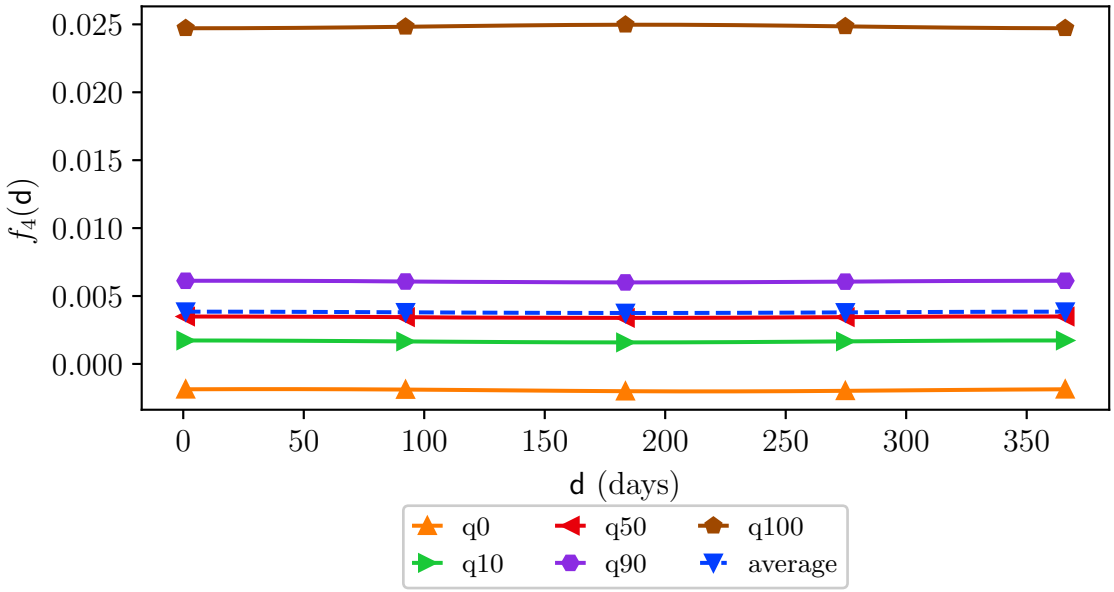


FIGURE F.35: Local effects of the day of the year
Quantiles of the univariate effect associated to the effect of the day of the year at each of the 32 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting models of the normalized loads. The regularization hyperparameter for the day of the year equals 100. It has been selected empirically and is quite large. The corresponding effects in the local models are consequently very limited.

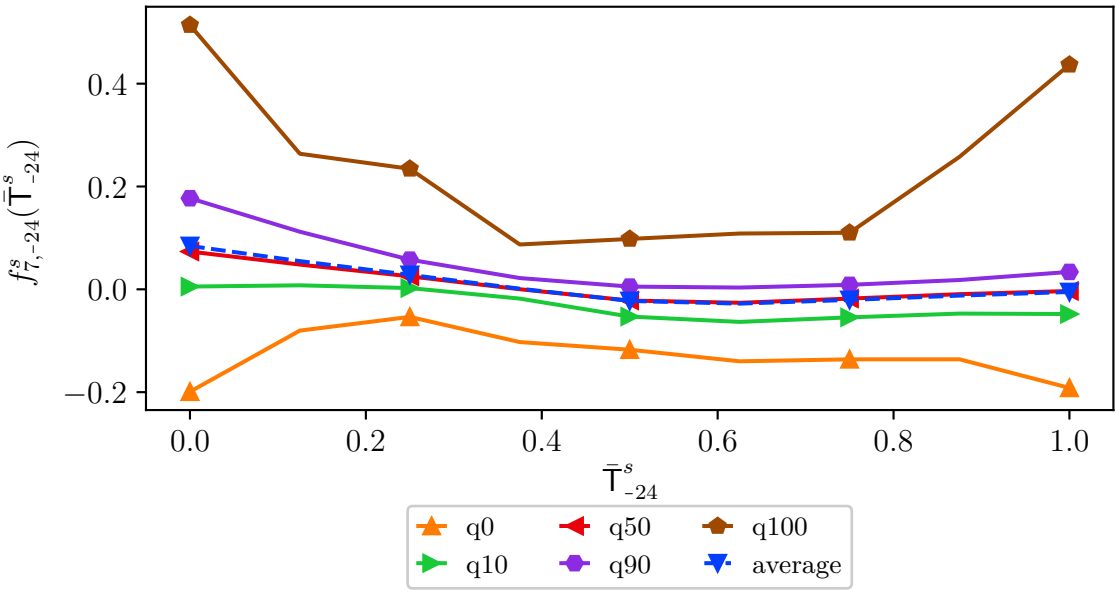


FIGURE F.36: Local effects of the 24 h max temperatures
Quantiles of the univariate effects associated to the maximum temperatures over the last 24 hours at each of the 9 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting models of the normalized loads.

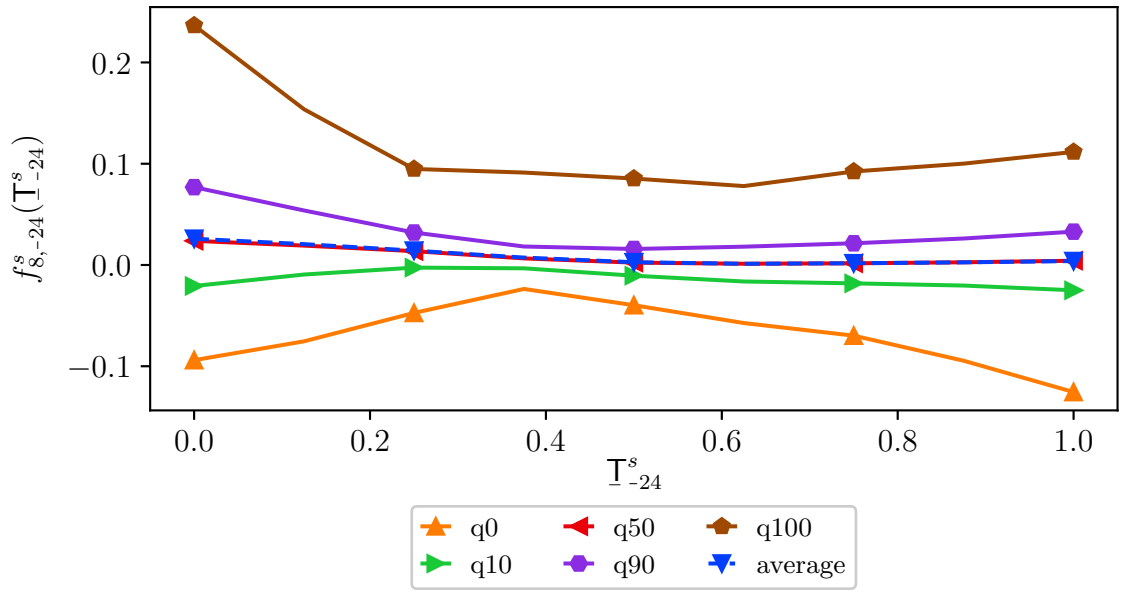


FIGURE F.37: Local effects of the 24 h min temperatures
Quantiles of the univariate effect associated to the minimum temperatures over the last 24 hours at each of the 9 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting models of the normalized loads.

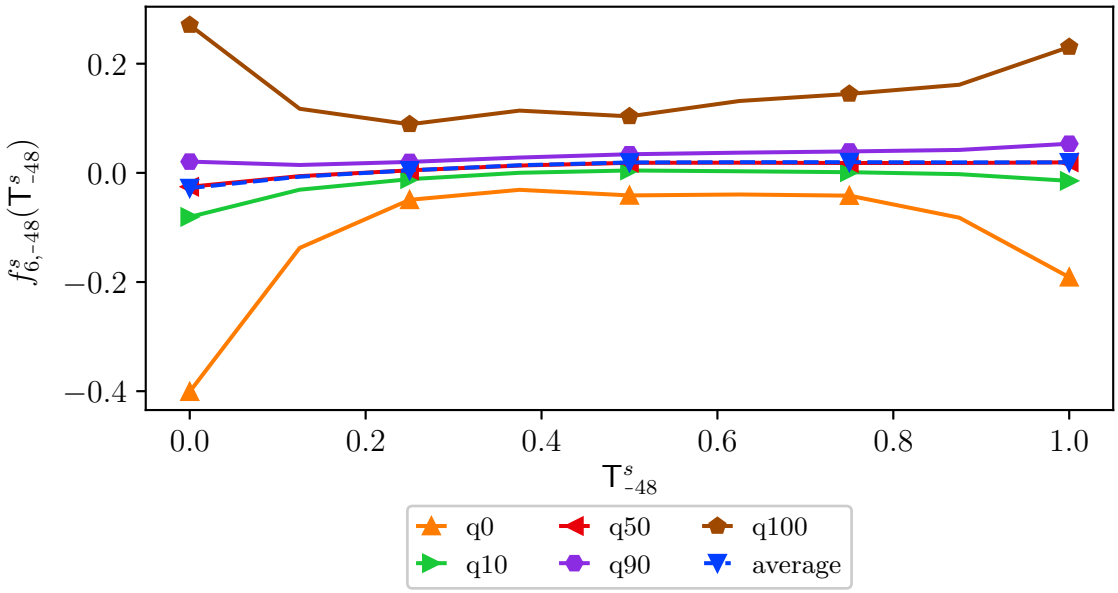


FIGURE F.38: Local effects of the 48 h-delayed temperatures
Quantiles of the univariate effects associated to the 48 hours-delayed temperatures at each of the 9 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting model of the normalized loads.

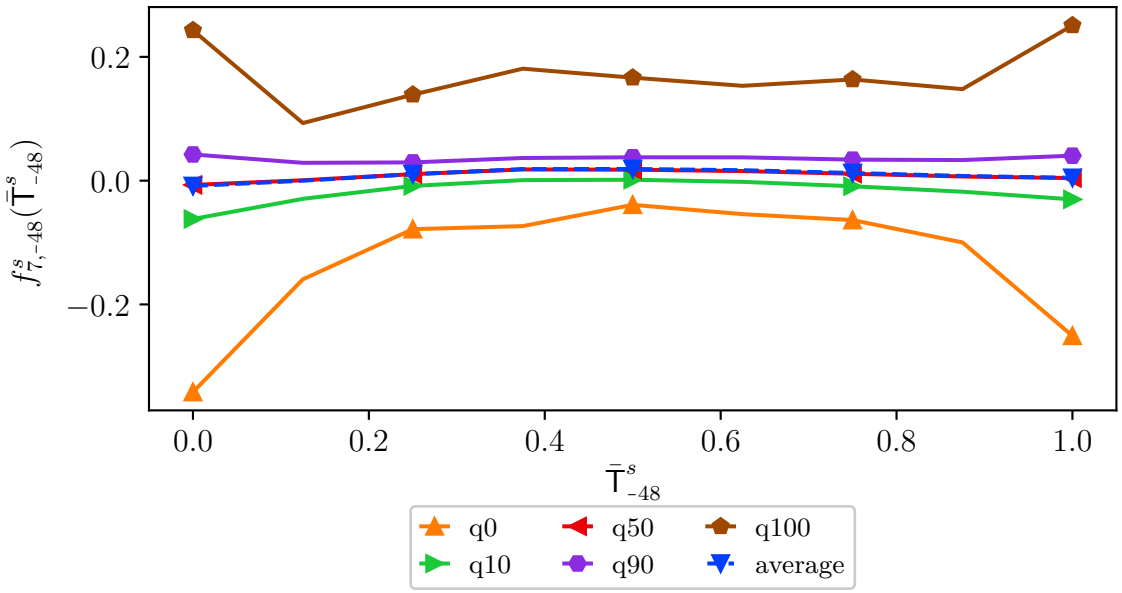


FIGURE F.39: Local effects of the 48 h max temperatures
Quantiles of the univariate effect associated to the maximum temperatures over the last 48 hours at each of the 9 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting models of the normalized loads.

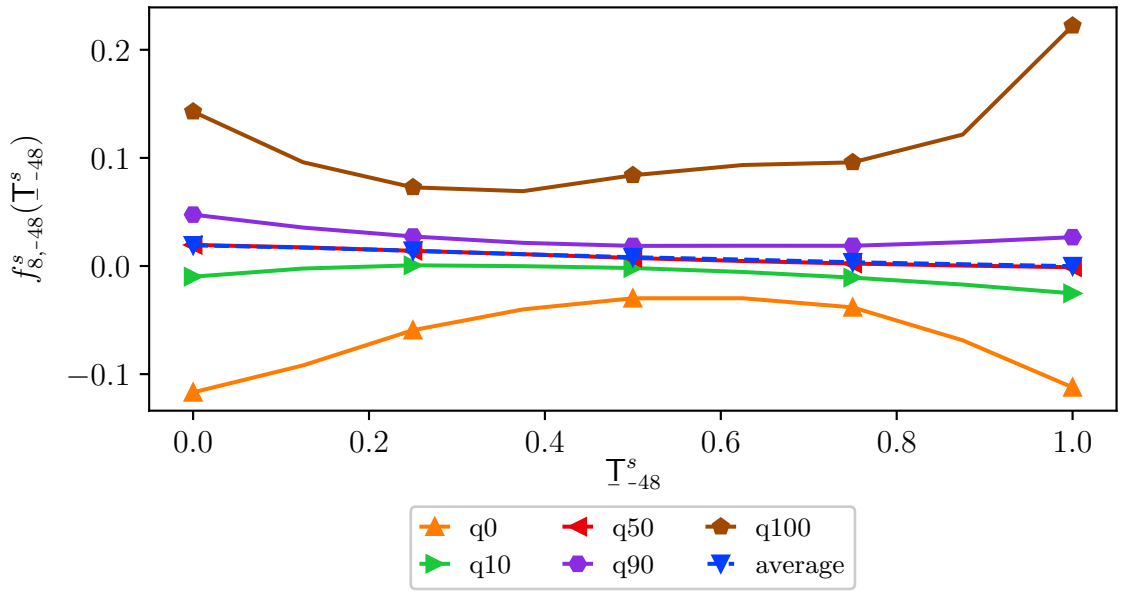


FIGURE F.40: Local effects of the 48 h min temperatures
Quantiles of the univariate effect associated to the minimum temperatures over the last 48 hours at each of the 9 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting models of the normalized loads.

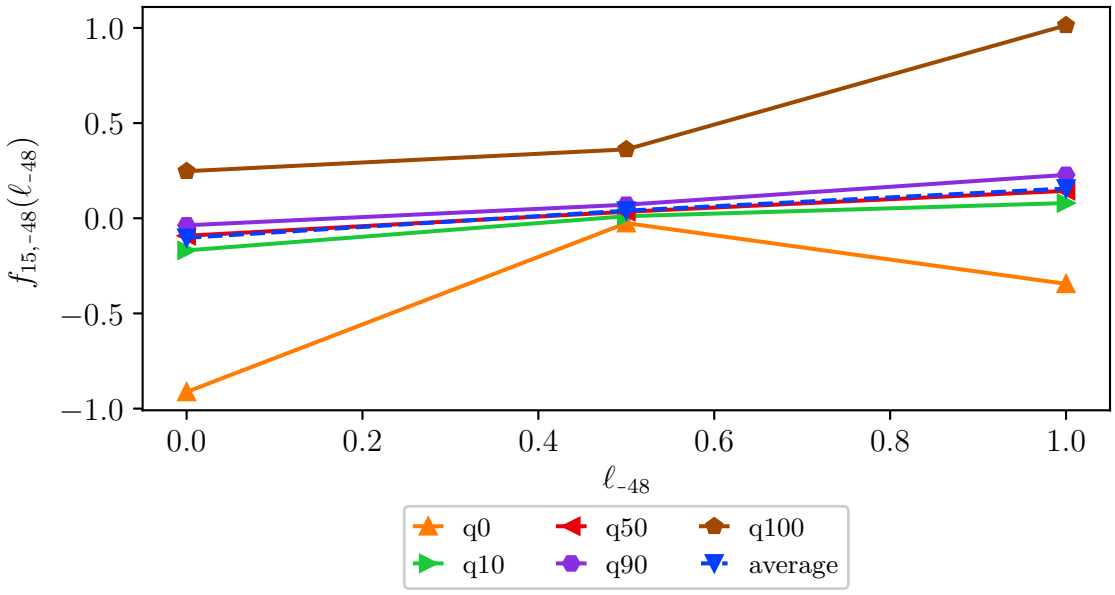


FIGURE F.41: Local effects of the 48 hours-delayed loads
Quantiles of the univariate effects associated to the 48 hours-delayed load at each of the 3 knots. Since the substations have different amplitudes, the coefficients correspond to the forecasting models of the normalized loads.

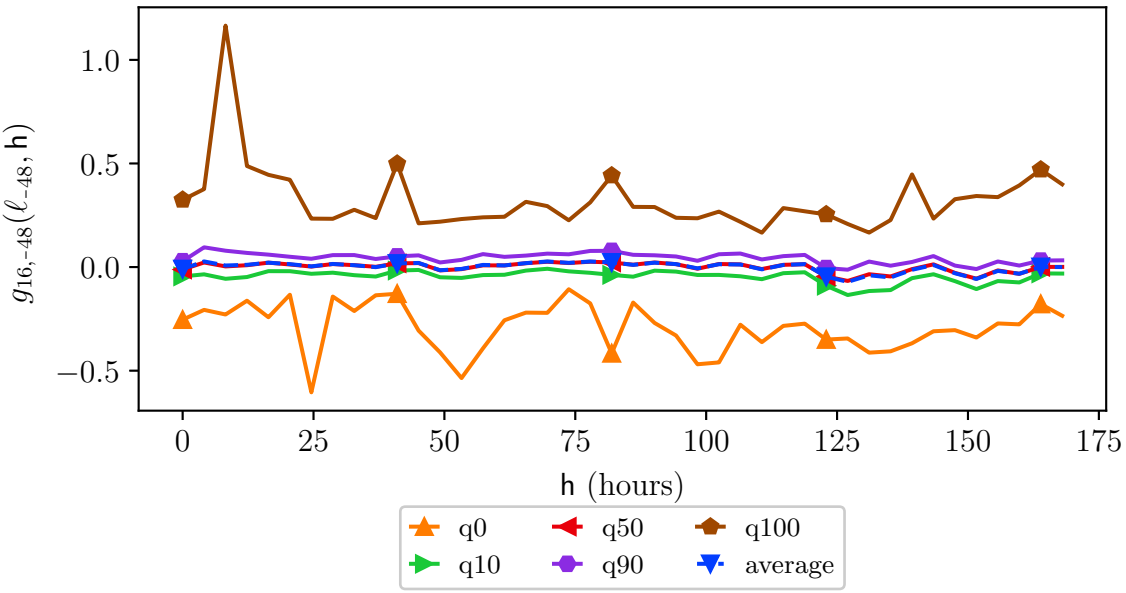


FIGURE F.42: Interactions of 48 h-delayed loads with hours
In the local model, the features for the past loads used to build the interaction with the hours of the week consist of a single linear function : there exists a function $h_{16,\text{-}48}$ such that $g_{16,\text{-}48}(\ell_{\text{-}48}, \mathsf{h}) = \ell_{\text{-}48} h_{16,\text{-}48}(\mathsf{h})$. The quantiles represented on the graph are the quantiles over the substations of the function $h_{16,\text{-}48}(\mathsf{h})$.

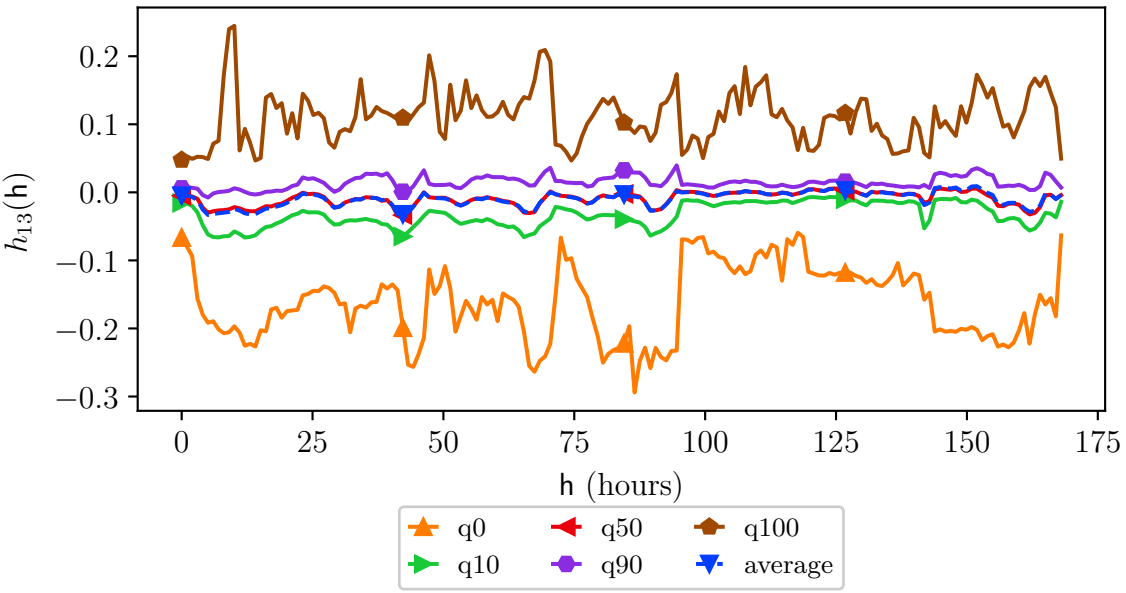


FIGURE F.43: Local effects of coming holidays
Quantiles of the local effects $h_{13}(\mathsf{h})$ of the coming holidays for the different hours of the week.

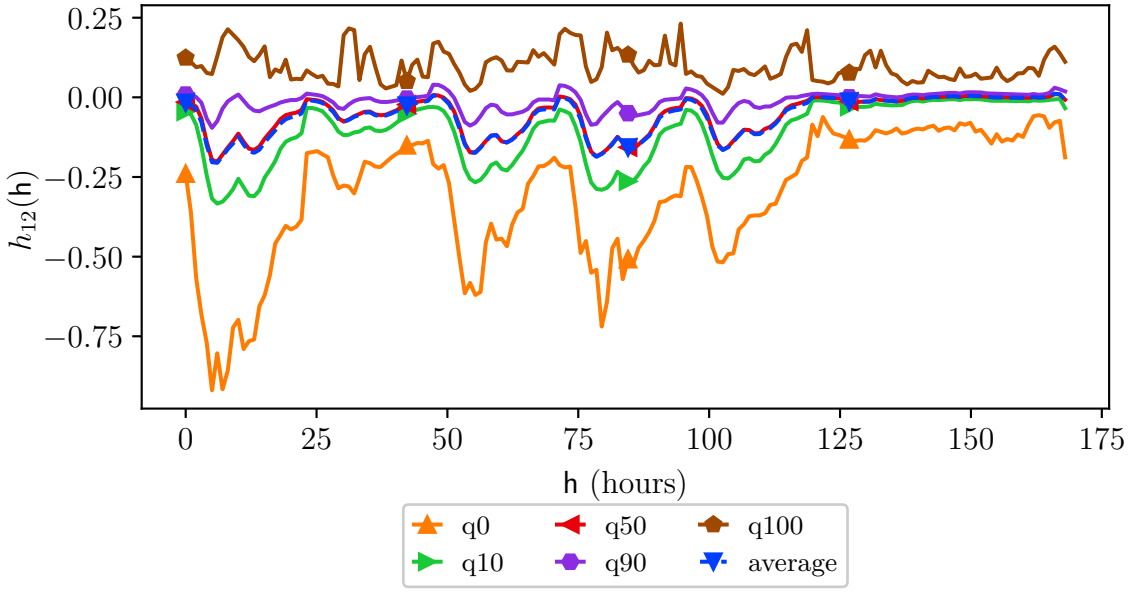


FIGURE F.44: Local effects of the holidays
Quantiles of the local effects $h_{12}(\mathsf{h})$ of the holidays for the different hours of the week.

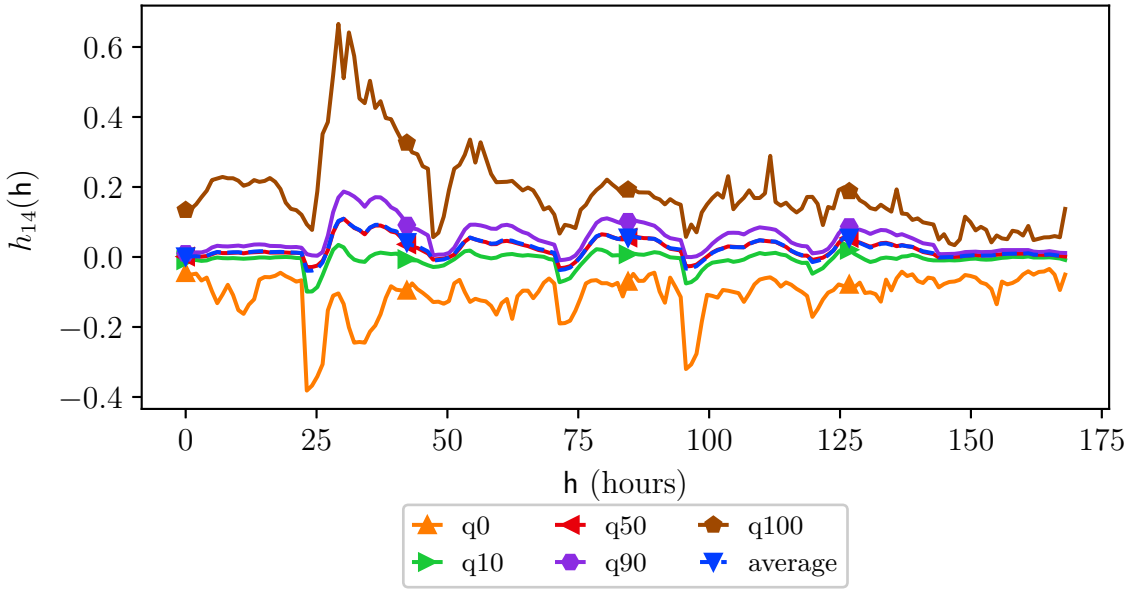


FIGURE F.45: Local effects of past holidays
Quantiles of the local effects $h_{14}(\mathsf{h})$ of the past holidays for the different hours of the week.

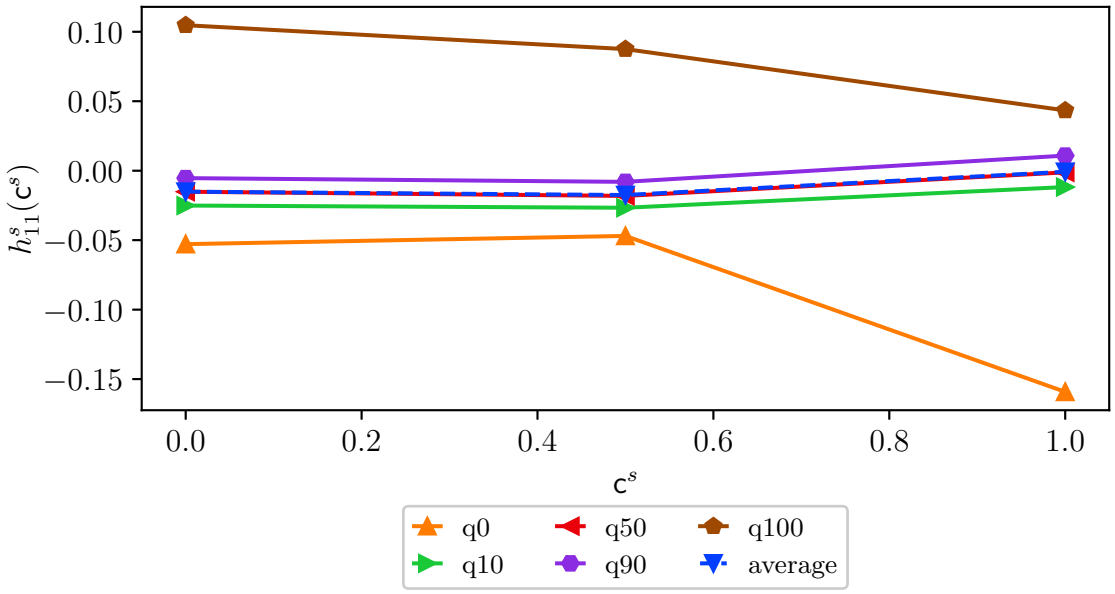


FIGURE F.46: Local effects of the cloud cover
Quantiles of the local effects $h^s_{11}(c^s)$ during the day of the cloud cover.

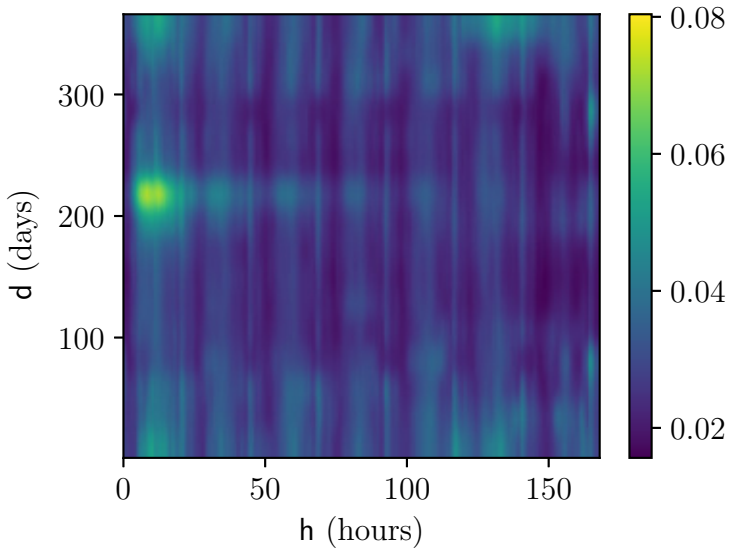


FIGURE F.47: Interactions of hours of the week with year days
Standard deviation over the substations of the interaction between the hour of the week and the day of the year.

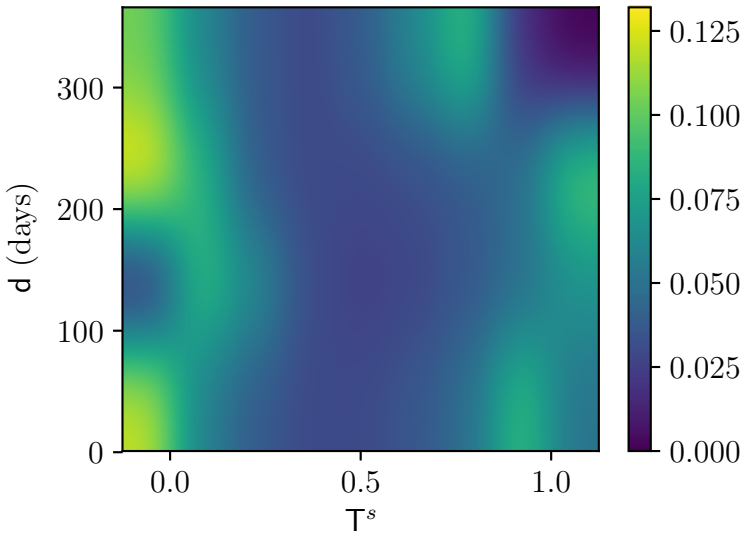


FIGURE F.48: Interactions between temperatures and year days
Standard deviation over the substations of the interaction between the temperature and the day of the year.

F.2.5 Evolution of the regularized univariate effects

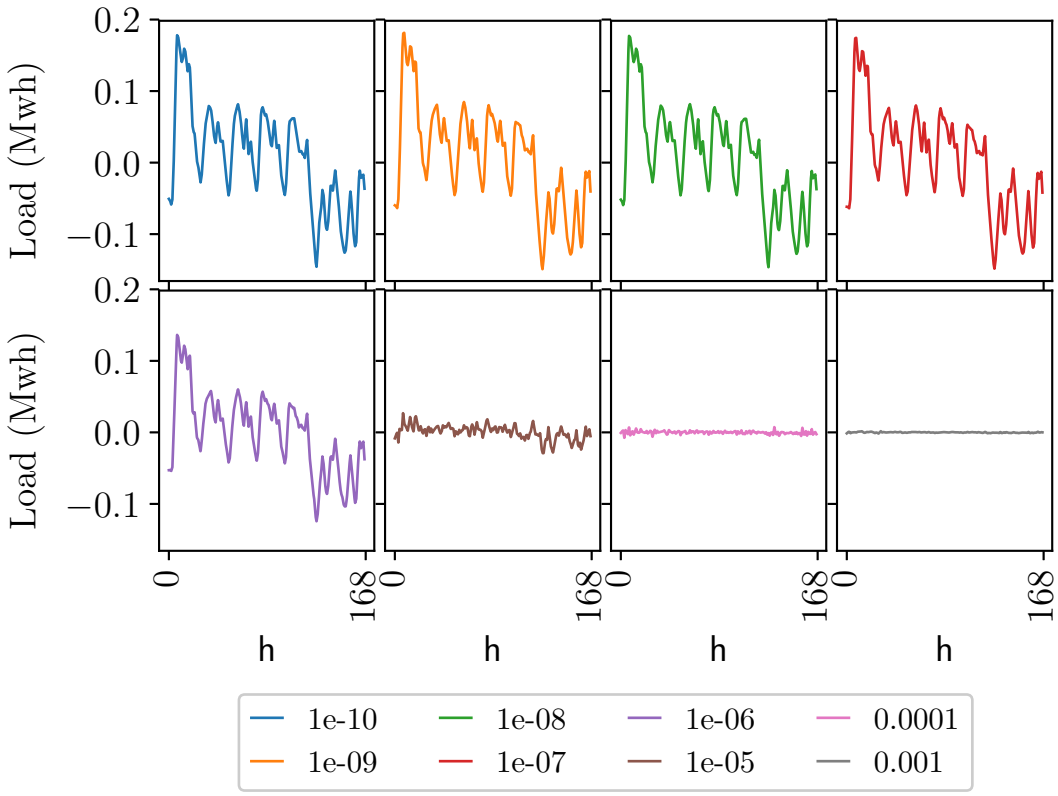


FIGURE F.49: Regularized effect of the hour of the week
Evolution of the univariate effect of the hour of the week in the national model, for different values of the regularization hyperparameter associated to the Ridge penalty. Empirically, we selected 10^{-8}.

F.2.6 Regularization of the national model

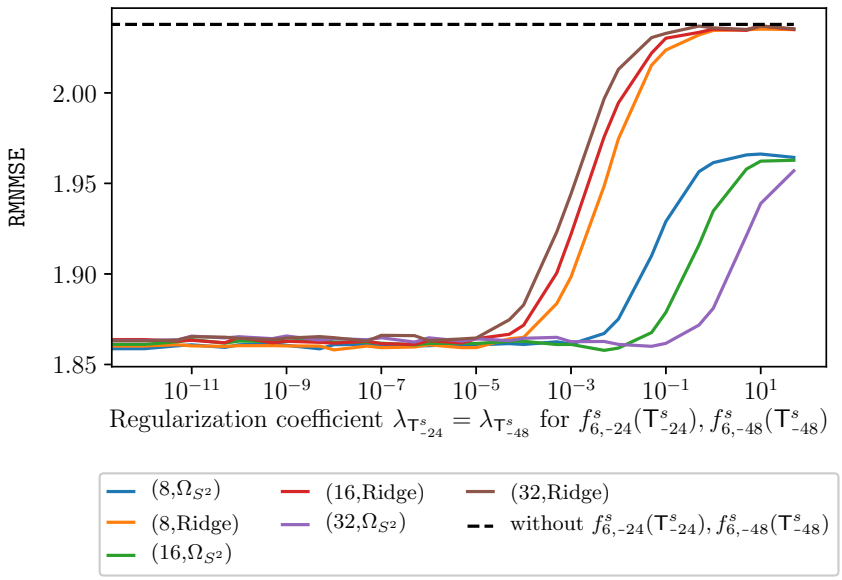


FIGURE F.50: Regularization of the 24 h-delayed temperature Performances with the national model on a test set of different number of knots for the 24 hours-delayed temperature univariate effect and two possible regularizations : Ridge and the Ω_{S^2} regularization.

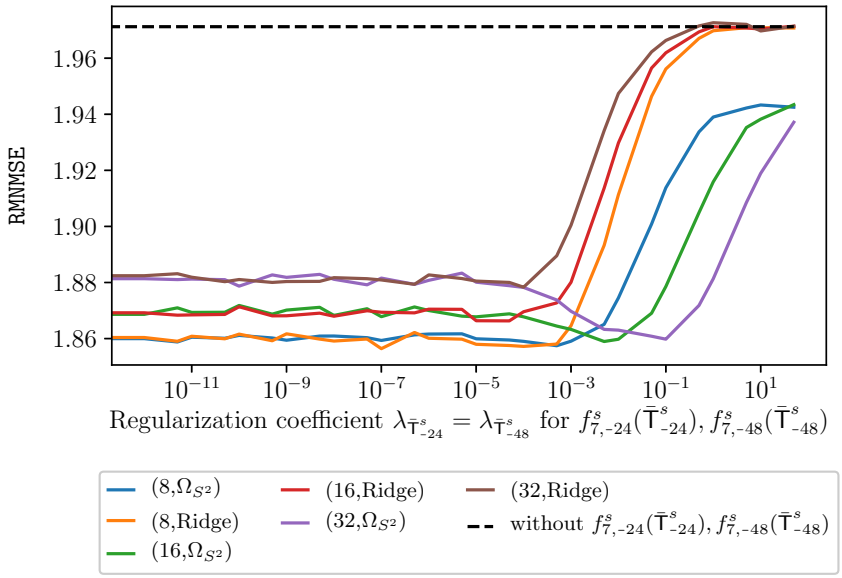


FIGURE F.51: Regularization of the past max temperatures
Performances with the national model on a test set of different number of knots for the univariate effect of the maximum temperatures and two possible regularizations : Ridge and the Ω_{S^2} regularization.

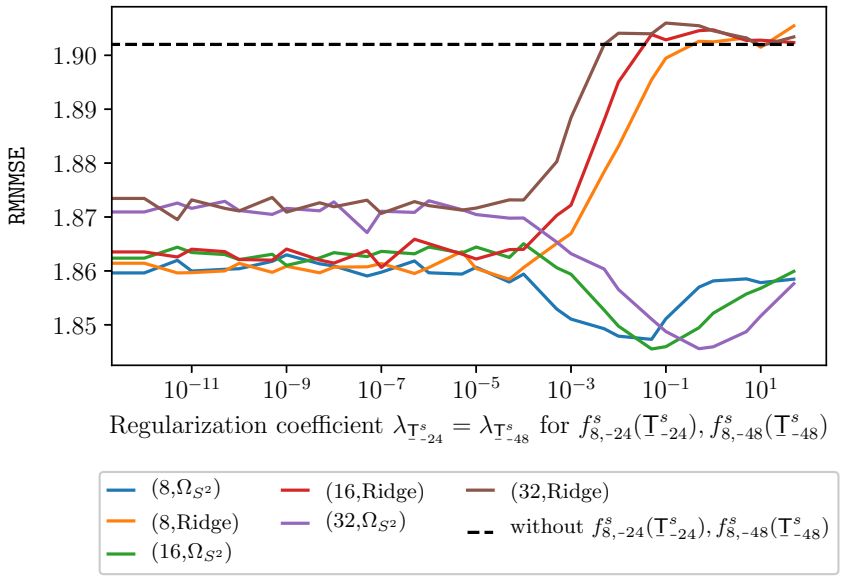


FIGURE F.52: Regularization of the past min temperatures
Performances with the national model on a test set of different number of knots for the univariate effect of the minimum temperatures and two possible regularizations : Ridge and the Ω_{S^2} regularization.

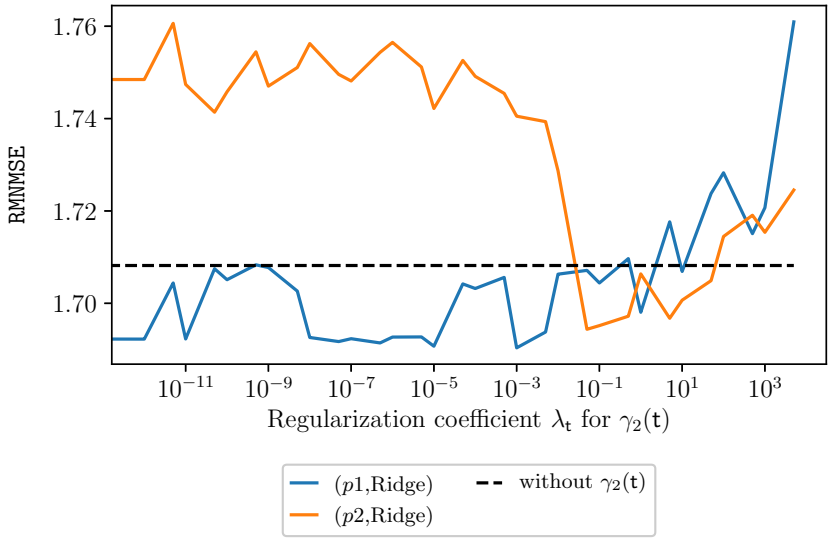


FIGURE F.53: Regularization of the effect of the timestamp
Performances with the national model on a test set of different parametrization of the timestamp univariate effect and with a Ridge regularization.

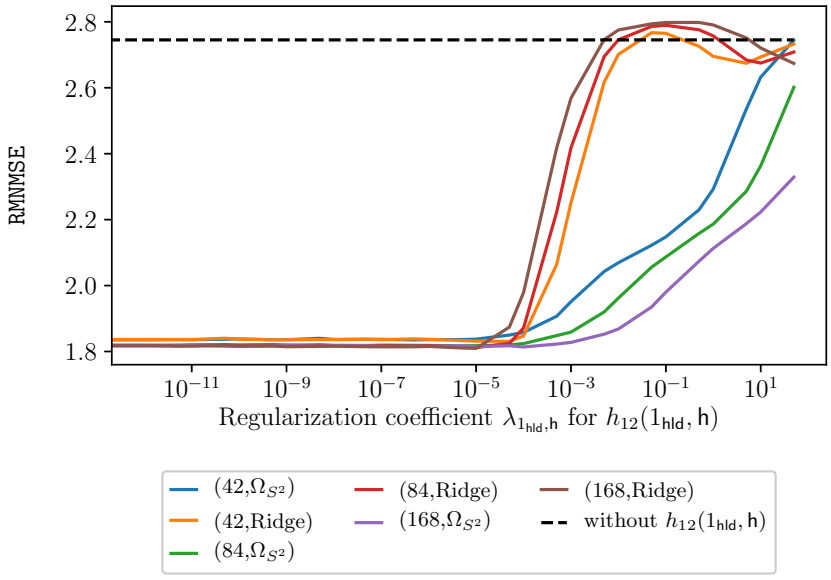


FIGURE F.54: Regularization of the holidays
Performances with the national model on a test set of different number of knots for the interactions between the indicator of holidays and the hour of the week with two possible regularizations : Ridge and the Ω_{S^2} regularization.

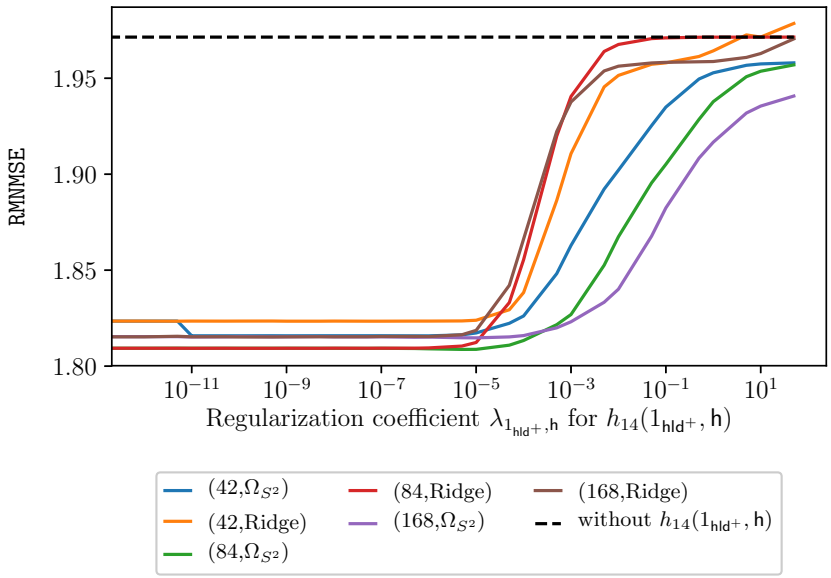


FIGURE F.55: Regularization of the coming holidays
Performances with the national model on a test set of different number of knots for the interactions between the indicator of days before a holiday and the hour of the week with two possible regularizations : Ridge and the Ω_{S^2} regularization.

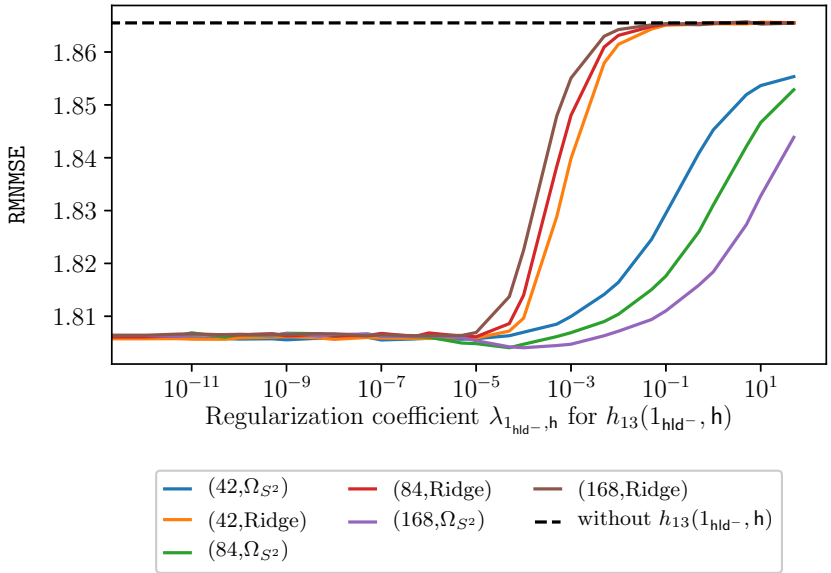


FIGURE F.56: Regularization of the past holidays
Performances with the national model on a test set of different number of knots for the interactions between the indicator of days before a holiday and the hour of the week with two possible regularizations : Ridge and the Ω_{S^2} regularization.

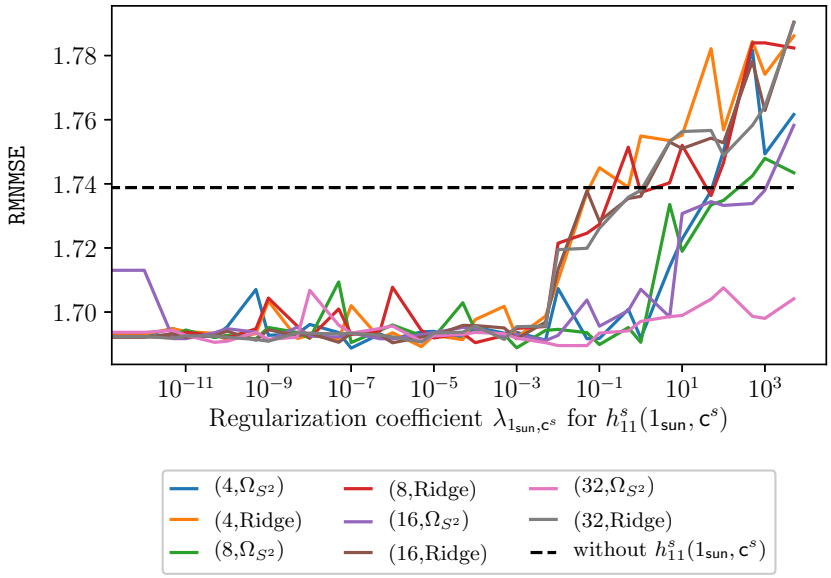


FIGURE F.57: Regularization of the cloud cover during the day
Performances with the national model on a test set of different number of knots for the interactions between the indicator of the daylight and the cloud cover with two possible regularizations : Ridge and the Ω_{S^2} regularization.

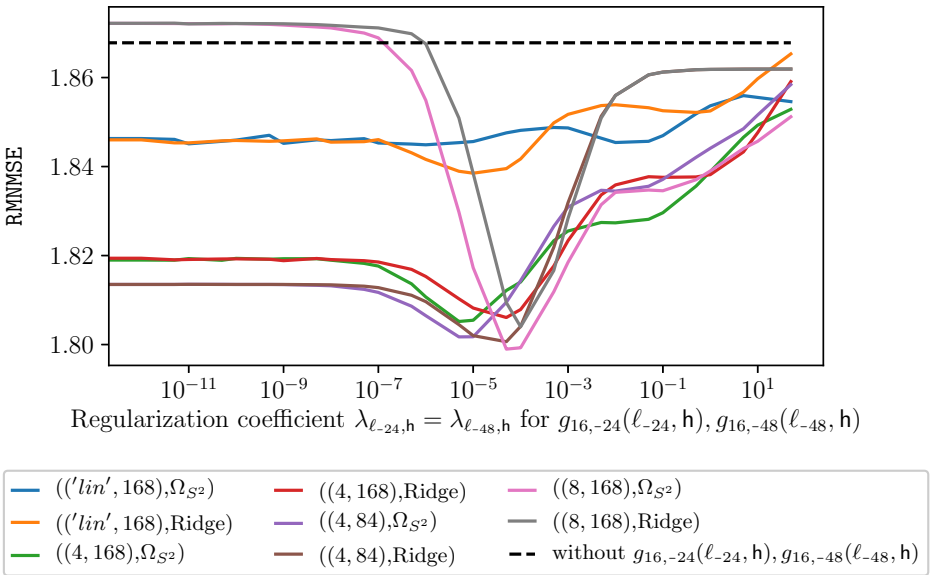


FIGURE F.58: Regularized past loads and hours of the week
Performances with the national model on a test set of different number of knots for the interactions between the past loads and the hour of the week with two possible regularizations : Ridge and the Ω_{S^2} regularization.

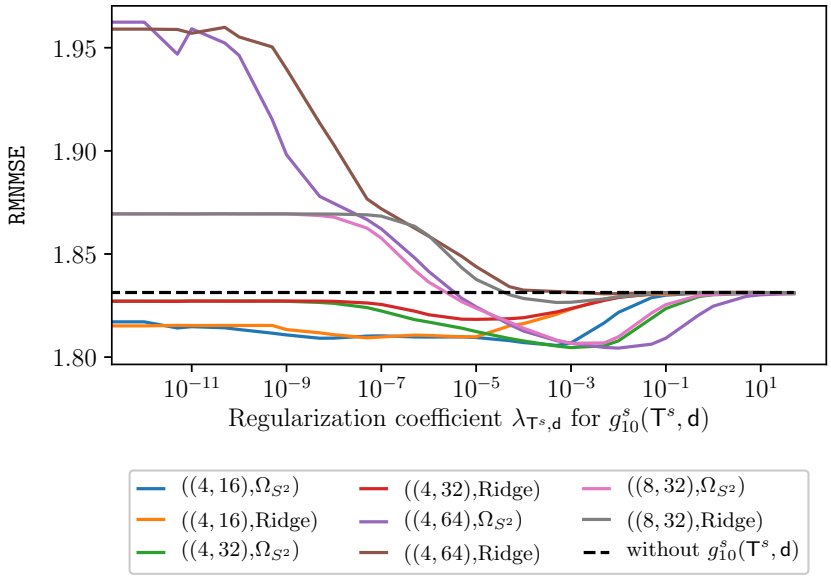


FIGURE F.59: Regularized temperatures and days of the year
Performances with the national model on a test set of different number of knots for the interactions between the temperature and the day of the year with two possible regularizations : Ridge and the Ω_{S^2} regularization.

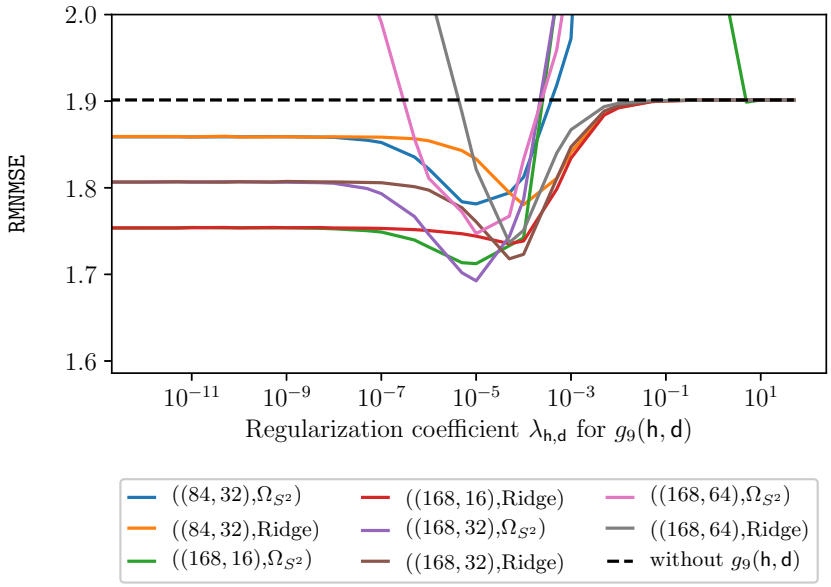


FIGURE F.60: Regularized hours and days of the year
Performances with the national model on a test set of different number of knots for the interactions between the hour of the week and the day of the year with two possible regularizations : Ridge and the Ω_{S^2} regularization.

F.2.7 Regularization of the local models

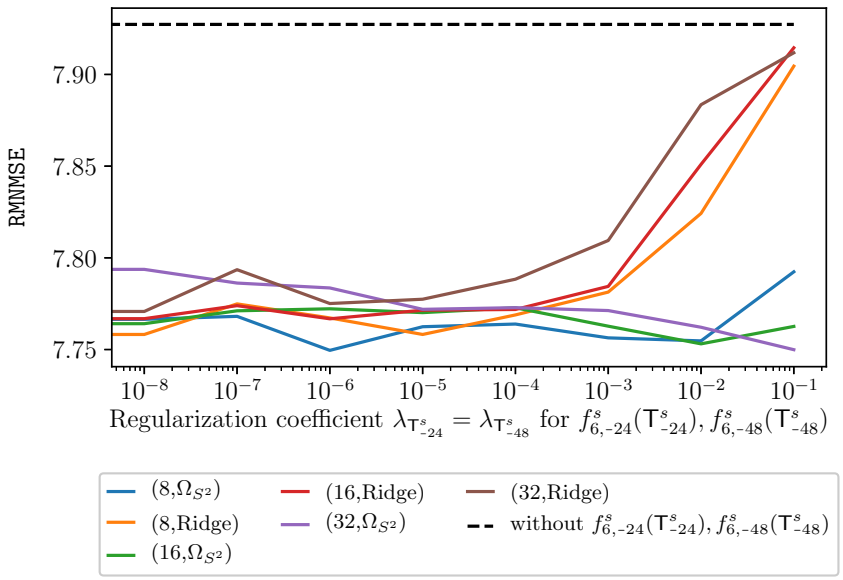


FIGURE F.61: Regularization of the effects of past temperatures
Performances with the local models on a test set of different number of knots for the univariate effect of the 24 hours-delayed temperature and two possible regularizations : Ridge and the Ω_{S^2} regularization.

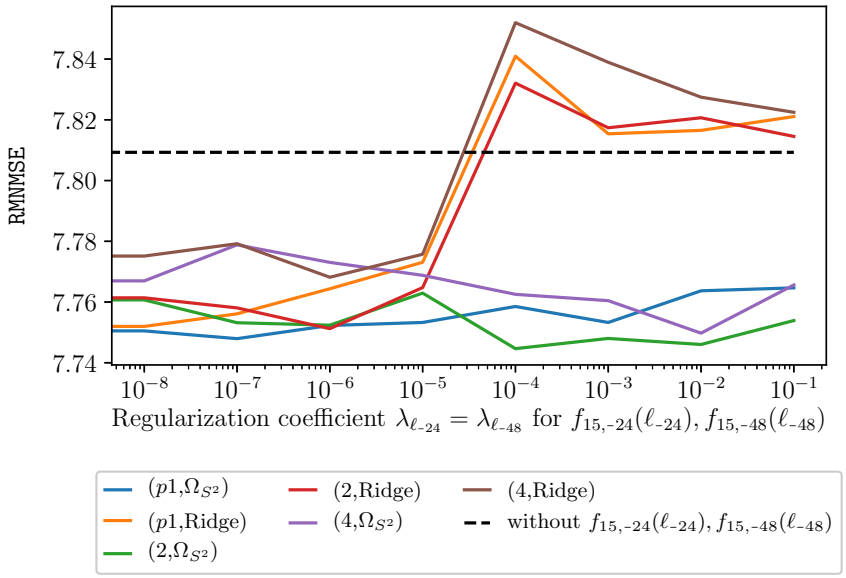


FIGURE F.62: Regularization of the local past loads
Performances with the local models on a test set of different number of knots for the univariate effect of the 24 hours-delayed load and two possible regularizations : Ridge and the Ω_{S^2} regularization.

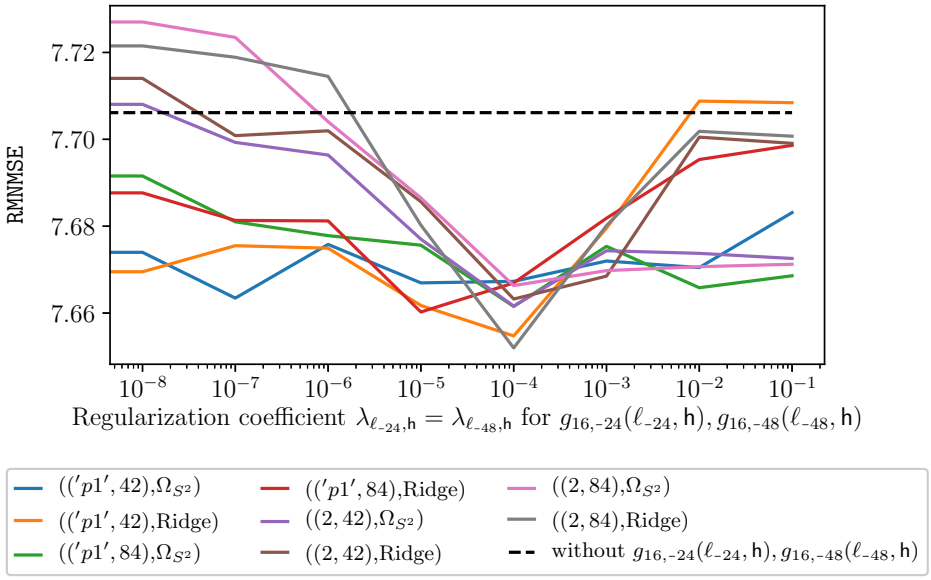


FIGURE F.63: Regularization of local past loads and hours
Performances with the local models on a test set of different number of knots for the interactions between the past loads and the hour of the week with two possible regularizations : Ridge and the Ω_{S^2} regularization.

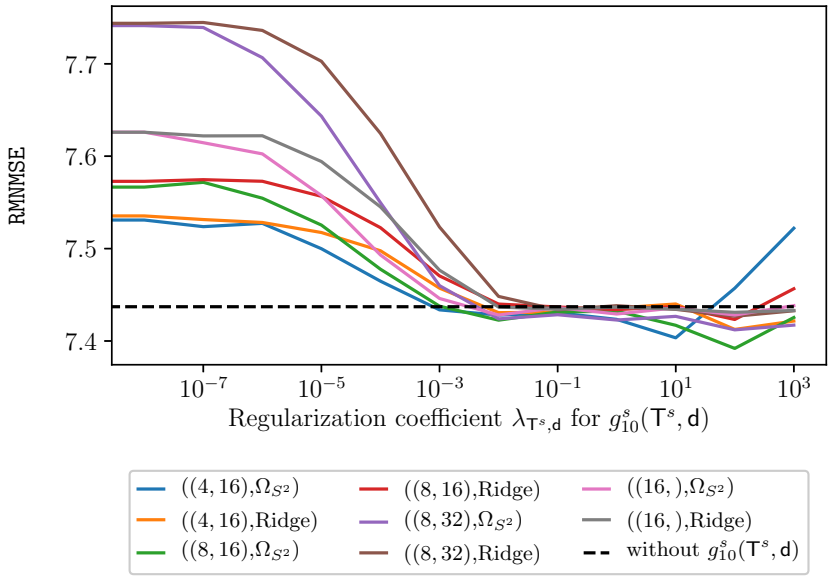


FIGURE F.64: Regularization of past temperatures and days
Performances with the local models on a test set of different number of knots for the interactions between the temperature and the day of the year with two possible regularizations : Ridge and the Ω_{S^2} regularization.

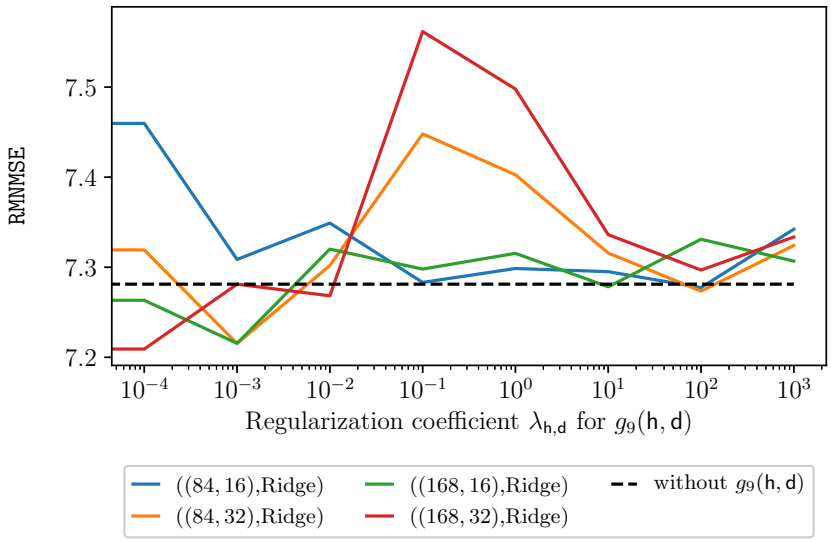


FIGURE F.65: Regularization of the hours and the year days Performances with the local models on a test set of different number of knots for the interactions between the hour of the week and the day of the year with two possible regularizations : Ridge and the Ω_{S^2} regularization.

F.2.8 Analysis of the temperatures

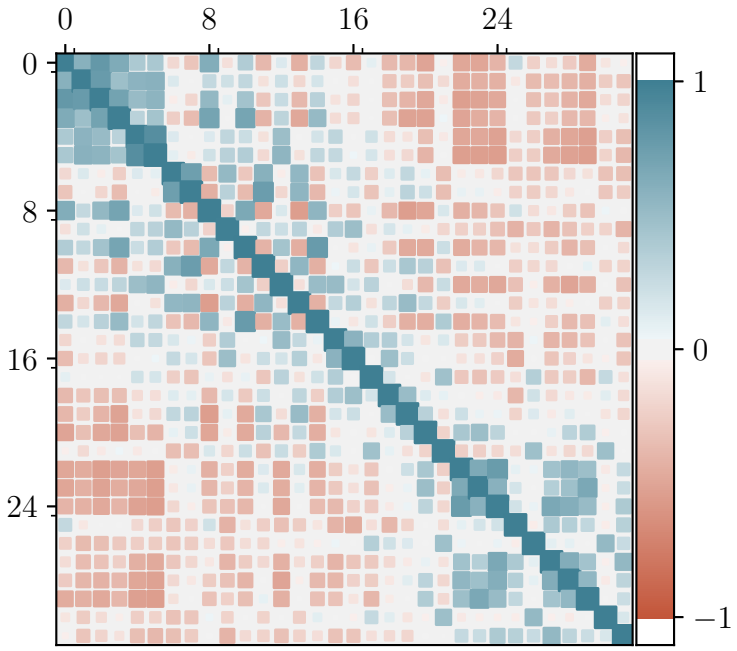


FIGURE F.66: Correlations of the 32 weather stations
Correlations between the columns of the matrix $\tilde{\boldsymbol{T}}$, obtained by centering the rows of the original temperature matrix $\boldsymbol{T} \in \mathbb{R}^{n,\mathcal{W}}$ containing the temperatures for the n observations during the 5 years in the dataset at the $\mathcal{W}$ different weather stations. The weather stations are sorted from North to South.

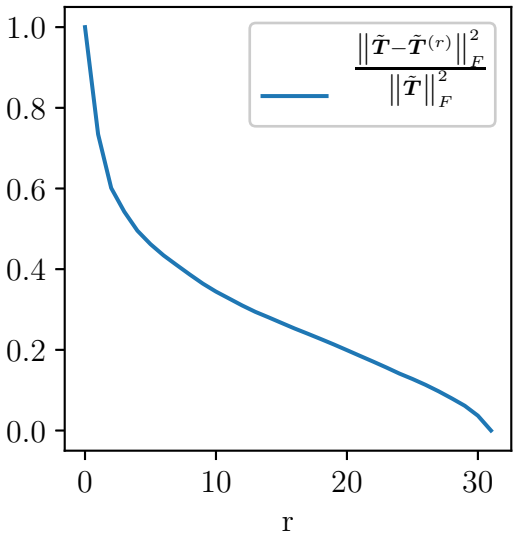


FIGURE F.67: Low-rank approximation of the temperatures
Norm of the residual matrix after subtracting the best rank-r approximation of the centered temperatures matrix $\boldsymbol{T}$. The matrix $\tilde{\boldsymbol{T}}$ is obtained by centering the rows of the original temperature matrix $\boldsymbol{T} \in \mathbb{R}^{n,\mathcal{W}}$ containing the temperatures for the 5 years in the dataset at the $\mathcal{W}$ different weather stations. For $r \in \mathbb{N}$, the matrix $\tilde{\boldsymbol{T}}^{(r)}$ is the closest rank-r approximation of $\tilde{\boldsymbol{T}}$ in terms of the Frobenius norm (See Figure 2.22 for supplementary details).

F.3 Additional Figures and Tables for Chapter 4

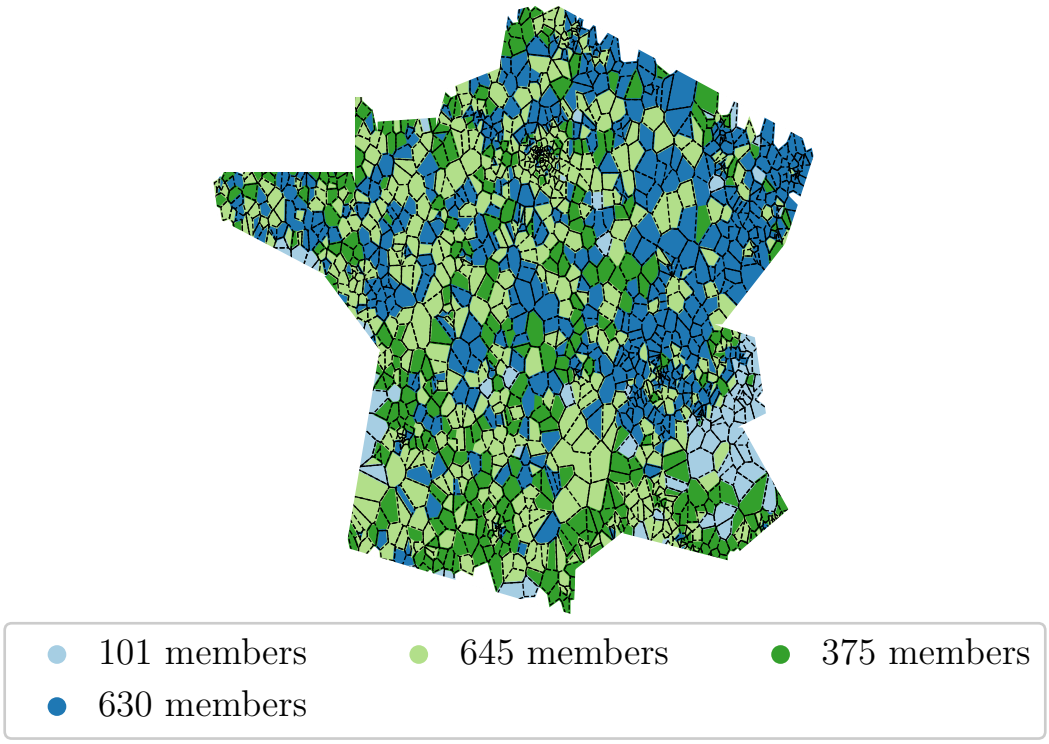


FIGURE F.68: Clustering of the coefficient vectors
The coefficient vectors estimated independently for the forecasting of the normalized loads of the different substations are clustered with a Kmeans algorithm.

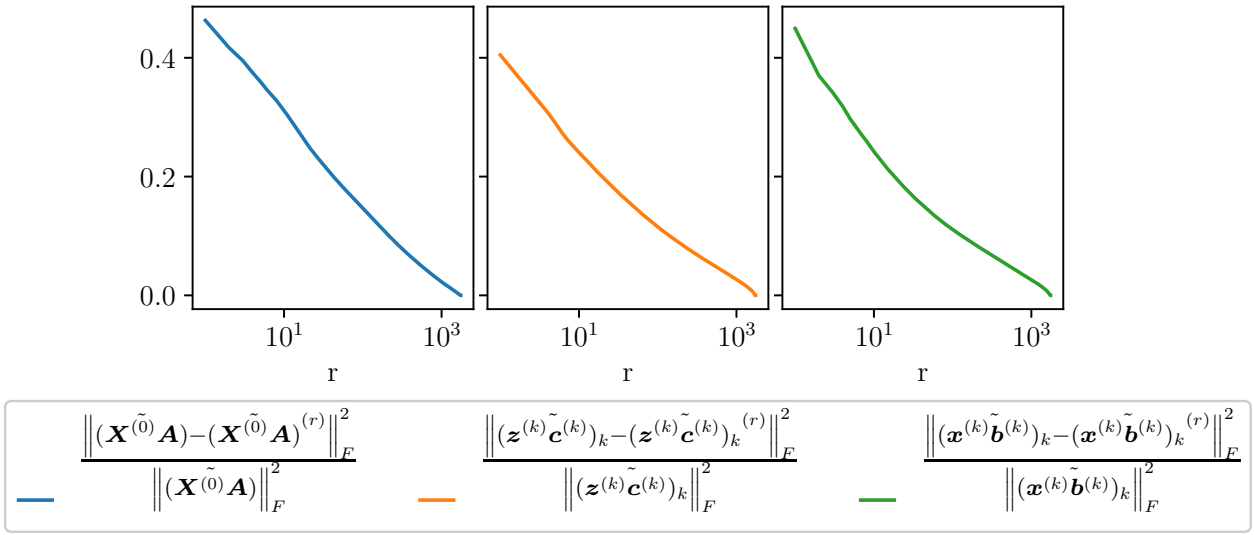


FIGURE F.69: Singular values of the prediction matrices
Norm of the residuals after subtracting the best rank-r approximation of the prediction matrices whose columns have been centered, from the common covariates, the individual covariates, and the sum of both.

Appendix G

Appendix to Chapter 5

G.1 Summary of results

We summarize the main steps of our paper, in red for RRR and in cyan for SRRR.

Corollary 13/14 : Strong convexity on cones of f/F^λ for RRR/SRRR	$\Rightarrow$	Corollary 16/17 : (proximal) PŁ inequality	$\Rightarrow$ with Theorem 15	Corollary 18 : Local linear convergence

We also summarize the different results obtained.

Results	RRR ($\lambda = 0$)		SRRR ($0 < \lambda$)	
Local minima are global minima	✓ Lemma 10		✗	
Algorithm	cst_st	ls	cst_st	ls
Global convergence to a critical point	✓ Theorem 48	✓ Theorem 48	$(*)$ Theorem 51	$(*)$ Theorem 51
Local linear convergence	✓ Corollary 18	✓ Corollary 18	✓($\lambda < \bar{\lambda}$) Corollary 18	✓($\lambda < \bar{\lambda}$) Corollary 18

- cst_st : Algorithm 1 with fixed step size $t \leq \frac{1}{L_X}$.
- ls : Algorithm 1 with line search.
- $(*)$: All limit points of the sequence are critical points. If these limit points are local minima and if for any $S \subset \{1, \dots, p\}$ of cardinality at least r, the matrix $X_S^T Y$ is full-rank, then Algorithm 1 converges to a local minimum (see Appendix G.6.2).

G.2 Additional definitions and classical results

In Sections G.2.1, G.2.2 and G.2.3, we give a few definitions that are used throughout the paper. We also recall classical results in Fact 20 and Fact 23. In Section G.2.4, we present the limiting subdifferential and a result for subanalytic functions in Lemma 28.

G.2.1 Strong convexity

Definition 19. *Given $d > 0$, $\mu > 0$ and a convex set $\mathcal{V} \subset \mathbb{R}^d$, a function $f : x \in \mathcal{V} \mapsto f(x)$ is μ-strongly convex if :*

$$\text{for all } x, y \in \mathcal{V},\ t \in [0,\, 1], \quad f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)-\frac{\mu}{2}t(1-t)\left\|y-x\right\|^2.$$

Fact 20. *Given $d > 0$, $\mu > 0$, a convex set $\mathcal{V} \subset \mathbb{R}^d$ and a differentiable function $f : x \in \mathcal{V} \mapsto f(x)$, f is μ-strongly convex if and only if :*

$$\text{for all } x, y \in \mathcal{V}, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\left\|y-x\right\|^2.$$

G.2.2 Smoothness and Lipschitz gradients

Definition 21. *Given $d > 0$, $L > 0$ and a set $\mathcal{V} \subset \mathbb{R}^d$, we say that a differentiable function $f : x \in \mathcal{V} \mapsto f(x)$ has L-Lipschitz gradients in $\mathcal{V}$ if :*

$$\text{for all } x, y \in \mathcal{V}, \quad \left\|\nabla f(x) - \nabla f(y)\right\| \leq L\left\|y-x\right\|.$$

Definition 22. *Given $d > 0$, $L > 0$ and a set $\mathcal{V} \subset \mathbb{R}^d$, we say that a function $f : x \in \mathcal{V} \mapsto f(x)$ is L-smooth in $\mathcal{V}$ if it is differentiable and such that :*

$$\text{for all } x, y \in \mathcal{V}, \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\left\|y-x\right\|^2.$$

Fact 23. *If f has L-Lipschitz gradients and $\mathcal{V}$ is convex, then f is L-smooth. If f is convex and L-smooth, then f has L-Lipschitz gradients.*

G.2.3 Sublevel sets

Definition 24. *Given a set $\mathcal{X}$ and a function $f : x \in \mathcal{X} \mapsto f(x)$, a set $\mathcal{V} \subset \mathcal{X}$ is called a sublevel set of the function f if there is $c \in \mathbb{R}$ such that :*

$$\mathcal{V} = \{x \in \mathcal{X},\, f(x) \leq c\}.$$

G.2.4 Subdifferentials, graph continuity and the KŁ property

Definition 25. *Given a real-valued extended function $F : \mathbb{R}^d \mapsto \mathbb{R} \cup \{\infty\}$, let*

$$dom\ F := \left\{x \in \mathbb{R}^d \mid F(x) < \infty\right\}$$

denote its domain. For each $x \in dom\ F$, the Fréchet subdifferential of F at x, written $\hat{\partial}F(x)$, is the set of vectors $v \in \mathbb{R}^d$ which satisfy :

$$\lim \inf_{y \neq x,\, y \to x} \frac{1}{\|y - x\|} \left[F(y) - F(x) - \langle v, y - x\rangle\right] \geq 0.$$

When $x \notin dom\ F$, we set $\hat{\partial}F(x) = \emptyset$. Given $x \in \mathbb{R}^d$, The limiting-subdifferential $\partial F(x)$ is defined as :

$$\partial F(x) := \left\{v \in \mathbb{R}^d \ \mid \ \exists\ x^k \to x,\ f(x^k) \to f(x),\ v^k \in \hat{\partial}F(x^k) \to v\right\},$$

$dom\ \partial F := \left\{x \in \mathbb{R}^d \mid \partial F(x) \neq \emptyset\right\}$ and the graph of ∂F is defined as :

$$graph(\partial F) := \left\{(x, u) \in \mathbb{R}^d \times \mathbb{R}^d \ \mid \ u \in \partial F(x)\right\}.$$

Fact 26. *[From Rockafellar and Wets, 2009] Let $F : \mathbb{R}^d \mapsto \mathbb{R}$ be a lower semi-continuous function and consider a sequence $\{(x_k, u_k)\}_{k\geq 0} \in graph(\partial F)^{\mathbb{N}}$ such that the sequence $\{(x_k, u_k, F(x_k))\}_{k\geq 0}$ converges to a point $\{(x, u, F(x))\}$. Then $(x, u) \in graph(\partial F)$.*

Definition 27. *[From Attouch et al., 2013] The function $F : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is said to have the Kurdyka-Łojasiewicz property at $x^* \in dom\ \partial F$ if there exists $\eta \in (0, +\infty]$, a neighborhood $\mathcal{U}$ of x^* and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that :*

1. *$\varphi(0) = 0$,*
2. *φ is $\mathcal{C}^1$ on $(0, \eta)$ and continuous at 0,*
3. *for all s in $(0, \eta)$, $\varphi'(s) > 0$,*
4. *for all $x \in \mathcal{U} \cap \{y \mid F(x^*) < F(y) < F(x^*) + \eta\}$, the Kurdyka-Łojasiewicz inequality holds*
$$\varphi'(F(x) - F(x^*))\ dist(0, \partial F(x)) \geq 1.$$

*Proper lower semi-continuous functions which satisfy the Kurdyka-Łojasiewicz inequality at each point of dom ∂F are called **KŁ** functions. Besides, KŁ with exponent α means the KŁ property with a function $\varphi : s \mapsto cs^{1-\alpha}$ where $c > 0$. We denote this property **KŁ-**α.*

Lemma 28. *[From Bolte et al., 2007] Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a subanalytic function with closed domain and assume that $F|_{dom\ F}$ is continuous. Then for any $x \in dom\ F$, there exist a neighborhood $\mathcal{V} \subset \mathbb{R}^d$ of x, an exponent $\theta \in [0, 1)$ and a constant $C > 0$ such that for all $y \in \mathcal{V}$, we have :*

$$|F(y) - F(x)|^\theta \leq C\, dist(0, \partial F(y)).$$

Note that norms and in particular the Frobenius norm, the trace-norm and the group-Lasso norm satisfy the KŁ property, so the functions that we consider in this paper satisfy this property.

G.2.5 Critical and KW-stationary points

Definition 29. *We say that $x \in \mathbb{R}^d$ is a critical point of F if $0 \in \partial F(x)$ where $\partial F(x)$ is defined in Definition 25.*

Definition 30. *Given a function $F := f_1 - f_2 + \lambda h$ where f_1 is differentiable while f_2 and h are proper, lower semi-continuous and convex, we say that $x \in \mathbb{R}^d$ is a KW-stationary point if there exist $u(x) \in \partial f_2(x)$ and $v(x) \in \partial h(x)$ such that :*

$$\nabla f_1(x) - u(x) + v(x) = 0.$$

Remark 31. *Note that the Definition 29 of critical points and the Definition 30 of KW-stationary points coincide when the function f_2 is differentiable.*

G.3 The Orthogonal Procrustes Problem

Given a matrix $M \in \mathbb{R}^{p,k}$ with $p \geq k$, we use at several points in the paper the following results that were presented in the Proof of Lemma 6 in [Ge et al., 2017].

Fact 32. *If $M = M_1^T M_2$, then*

$$\max_{V \in \mathbb{R}^{p,k} :\, V^T V = I_k} \langle M, V \rangle \text{ has the same set of optima as } \min_{V \in \mathbb{R}^{p,k} :\, V^T V = I_k} \frac{1}{2} \|M_2 - M_1 V\|_F^2 .$$

Fact 33. *The optimal value of the following orthogonal Procrustes problem is given by :*

$$\max_{V \in \mathbb{R}^{p,k} :\, V^T V = I_k} \langle M, V \rangle = \|M\|_* .$$

Fact 34. *If $R_1 \Sigma R_2^T$ is a complete singular value decomposition of M where $R_1 \in \mathbb{R}^{p,p}$ is such that $R_1^T R_1 = I_p$, $\Sigma \in \mathbb{R}_+^{p,k}$ has non-zero elements $\sigma_1 \geq \ldots \geq \sigma_k \geq 0$ only on the diagonal and $R_2 \in \mathbb{R}^{k,k}$ is such that $R_2^T R_2 = I_k$, then an optimal solution of the orthogonal Procrustes problem is given by*

$$R_1 \begin{bmatrix} I_k \\ 0_{p-k,k} \end{bmatrix} R_2^T \quad \in \quad \operatorname*{argmax}_{V \in \mathbb{R}^{p,k} :\, V^T V = I_k} \langle M, V \rangle.$$

Fact 35. *With the same notations as in Fact 34, if M is full-rank then, although R_1 and R_2 are not uniquely defined, the following Procrustes problem has a unique solution :*

$$\operatorname*{argmax}_{V \in \mathbb{R}^{p,k} :\, V^T V = I_k} \langle M, V \rangle \quad = \quad \left\{ R_1 \begin{bmatrix} I_k \\ 0_{p-k,k} \end{bmatrix} R_2^T \right\}.$$

Fact 36. *If $p = k$ then $I_r \in \operatorname{argmax}_{V \in \mathbb{R}^{p,p} :\, V^T V = I_p} \langle M, V \rangle$ if and only if M is positive-semidefinite.*

Proof. Fact 32 comes by seeing that for any $V \in \mathbb{R}^{p,k}$ such that $V^T V = I_k$, we have :

$$\begin{aligned} \frac{1}{2} \|M_2 - M_1 V\|_F^2 &= \frac{1}{2} \|M_2\|_F^2 + \frac{1}{2} \|M_1 V\|_F^2 - 2\langle M_2, M_1 V \rangle \\ &= \frac{1}{2} \|M_2\|_F^2 + \frac{1}{2} \|M_1\|_F^2 - 2\langle M_1^T M_2, V \rangle. \end{aligned}$$

To prove Fact 33 and Fact 34, let $R_1 \Sigma R_2^T$ be a singular value decomposition of M where $R_1 \in \mathbb{R}^{p,k}$ is such that $R_1^T R_1 = I_k$, $\Sigma \in \mathbb{R}_+^{k,k}$ has nonzero elements $\sigma_1 \geq \ldots \geq \sigma_k \geq 0$ only on the diagonal and $R_2 \in \mathbb{R}^{k,k}$ is such that $R_2^T R_2 = I_k$. Also, let $R_1^\perp \in \mathbb{R}^{p,p-k}$ such that $R := \begin{bmatrix} R_1 & R_1^\perp \end{bmatrix}$ satisfies $R^T R = I_p$. Writing $M = R_1 \Sigma R_2^T$ and using the change of variables $V = R_1 A R_2^T + R_1^\perp B R_2^T$, we have :

$$\begin{aligned} & \max_{V \in \mathbb{R}^{p,k} \colon V^T V = I_k} \langle M, V \rangle \\ = & \max_{A \in \mathbb{R}^{k,k},\, B \in \mathbb{R}^{p-k,k} \colon A^T A + B^T B = I_k} \langle R_1 \Sigma R_2^T, R_1 A R_2^T + R_1^\perp B R_2^T \rangle \\ = & \max_{A \in \mathbb{R}^{p,p},\, B \in \mathbb{R}^{p-k,k} \colon A^T A + B^T B = I_k} \langle \begin{bmatrix} \Sigma \\ 0_{p-k,k} \end{bmatrix}, \begin{bmatrix} A \\ B \end{bmatrix} \rangle \\ = & \max_{C \in \mathbb{R}^{p,k} \colon C^T C = I_k} \langle \begin{bmatrix} \Sigma \\ 0_{p-k,k} \end{bmatrix}, C \rangle. \end{aligned} \tag{G.1}$$

Let $C \in \mathbb{R}^{p,k}$ such that $C^T C = I_k$, we have :

$$\begin{aligned} \langle \begin{bmatrix} \Sigma \\ 0_{p-k,k} \end{bmatrix}, C \rangle &= \sum_{i=1}^{k} \sigma_i C_{i,i} \\ &\leq \sum_{i=1}^{k} \sigma_i \\ &= \|\Sigma\|_* \\ &= \|M\|_* . \end{aligned} \tag{G.2}$$

We have Inequality (G.2) since Σ has only nonnegative coefficients and the columns of C have unit norm so $C_{i,i} \leq 1$ for all $1 \leq i \leq k$. Besides, Inequality (G.2) is attained for $C = \begin{bmatrix} I_k \\ 0_{p-k,k} \end{bmatrix}$ which corresponds in Problem (G.1) to $V = R_1 \begin{bmatrix} I_k \\ 0_{p-k,k} \end{bmatrix} R_2^T$. This proves Fact 33 and Fact 34.

To prove Fact 35, that is to say that $\operatorname{argmax}_{V \in \mathbb{R}^{p,k} \colon V^T V = I_k} \langle M, V \rangle$ is a singleton if M is full-rank, it is sufficient to notice that Inequality (G.2) is strict if all the σ_i are non-zero and $C_{i,i} \neq 1$ for some $1 \leq i \leq k$.

To prove Fact 36, note that $I_r \in \operatorname{argmax}_{V \in \mathbb{R}^{p,p} \colon V^T V = I_p} \langle M, V \rangle$ implies $\operatorname{tr}(M) = \|M\|_*$ with Fact 33 and this is only true for positive-semidefinite matrices. Conversely, if M is positive-semidefinite, then by Fact 34, we have :

$$I_r \in \operatorname*{argmax}_{V \in \mathbb{R}^{p,p} \colon V^T V = I_p} \langle M, V \rangle.$$

□

G.4 The Forward-Backward Descent Algorithm 1

Given $U \in \mathbb{R}^{p,r}$, we recall that we compute the forward direction for Algorithm 1 with the gradient $X^T X U$ of $U' \mapsto \frac{1}{2} \|XU'\|_F^2$ and z_U a subgradient of $U' \mapsto \left\|Y^T X U'\right\|_*$ whose computation is detailed in Appendix G.4.1.2. Setting with

a slight abuse of notation $\nabla f(U) := X^T XU - z_U$, then t and U_+ are obtained with Algorithm 2 such that the (LS) condition $\tilde{F}^{\lambda}_{t,U}(U^+) \geq F^{\lambda}(U^+)$ is satisfied where :

$$U_+ = \underset{U' \in \mathbb{R}^{p,r}}{\operatorname{argmin}} f(U) + \langle \nabla f(U), U' - U \rangle + \frac{1}{2t} \|U' - U\|_F^2 + \lambda \|U'\|_{1,2} \,. \tag{G.3}$$

G.4.1 Subgradients for the descent direction

If we strictly applied the subgradient-type algorithm proposed by Khamaru and Wainwright [2018] and computed a forward direction for Algorithm 1 with Fact 37, we could only prove global convergence to a KW-stationary point. Instead, we introduce in Appendix G.4.1.2 an additional condition on the subgradient that is leveraged in Appendix G.6 to guarantee convergence to a critical point.

G.4.1.1 Subgradients of $U \mapsto \left\|Y^T XU\right\|_*$.

Thanks to Fact 33 and Fact 34, we can easily compute subgradients of $f_2 : U \mapsto \left\|Y^T XU\right\|_*$.

Fact 37. *Let n, $p \geq 0$, $r \leq min(n, p)$, $X \in \mathbb{R}^{n,p}$, $Y \in \mathbb{R}^{n,k}$, $U \in \mathbb{R}^{p,r}$ and $R_1 D R_2^T$ be a singular value decomposition of $Y^T XU$ with $R_1 \in \mathbb{R}^{k,r}$, $R_1^T R_1 = I_r$, $D \in \mathbb{R}^{r,r}$ a diagonal matrix with nonnegative coefficients, $R_2 \in \mathbb{R}^{r,r}$ and $R_2^T R_2 = I_r$. We denote $V = R_1 R_2^T \in \mathbb{R}^{k,r}$. For any $U' \in \mathbb{R}^{p,r}$, we have :*

$$\left\|Y^T XU'\right\|_* \geq \left\|Y^T XU\right\|_* + \langle X^T YV, U' - U \rangle.$$

Therefore, $X^T YV$ is a subgradient of $f_2 : U' \mapsto \left\|Y^T XU'\right\|_*$ at U.

Proof. Let $U \in \mathbb{R}^{p,r}$ and $V \in \mathbb{R}^{k,r}$ be defined as in Fact 37. Since $V^T V = R_2 R_1^T R_1 R_2^T = I_r$, we have by Fact 33 and Fact 34 :

$$\left\|Y^T XU\right\|_* = \langle V, Y^T XU \rangle. \tag{G.4}$$

By Fact 33, we also have for any $U' \in \mathbb{R}^{p,r}$,

$$\left\|Y^T XU'\right\|_* \geq \langle V, Y^T XU' \rangle. \tag{G.5}$$

Combining Equation (G.4) and Equation (G.5), we obtain :

$$\left\|Y^T XU'\right\|_* \geq \left\|Y^T XU\right\|_* + \langle V, Y^T X(U' - U) \rangle.$$

□

Remark 38. *We could also obtain subgradients of f_2 using Danskin's Theorem [Danskin, 1967] but the proposed analysis in the proof of Fact 37 seems more explicit. Besides, the choice of a specific subgradient in Lemma 40 is pivotal for the global convergence analysis in Appendix G.6, as explained in Remark 39.*

G.4.1.2 Computations of z_U for Algorithm 1.

Here, we present how, given $U \in \mathbb{R}^{p,r}$, the subgradient of $f_2 : U \mapsto \left\|Y^TXU\right\|_*$ is built for Algorithm 1 and we do not assume necessarily that X^TX is full-rank. Therefore, we denote $(X^TX)^{\frac{\dagger}{2}}$ a square-root of the pseudo-inverse of X^TX and, PSQ^T the reduced singular value decomposition of $(X^TX)^{\frac{\dagger}{2}}X^TY$. If the latter has rank ℓ then $P \in \mathbb{R}^{p,\ell}$ and $Q \in \mathbb{R}^{k,\ell}$ have orthonormal columns and $S \in \mathbb{R}^{\ell,\ell}$ is the diagonal matrix with singular values $s_1 \geq \ldots \geq s_\ell > 0$. We also denote $M \in \mathbb{R}^{k,r}$ a matrix whose columns are orthonormal and belong to $\text{Im } Y^TX(X^TX)^{\frac{\dagger}{2}}$, we compute this matrix only once at the beginning of Algorithm 1 with a Gram-Scmidt process. When X^TX is invertible, the computational cost is significantly reduced since we then have $\text{Im } Y^TX(X^TX)^{\frac{\dagger}{2}} = \text{Im } Y^TX$.

To compute z_U for Algorithm 1 - given $U \in \mathbb{R}^{p,r}$ - we first compute a singular value decomposition LDR_2^T of Y^TXU with $c = \text{rank}(Y^TXU)$, $L \in \mathbb{R}^{k,r}$, $L^TL = I_r$, $D \in \mathbb{R}^{r,r}$ a diagonal matrix with nonnegative coefficients, $R_2 \in \mathbb{R}^{r,r}$ and $R_2^TR_2 = I_r$. The computational cost is $O(kr^2)$ and we write :

$$\begin{aligned} L &= \begin{bmatrix} L^{>0} & L^0 \end{bmatrix}, \quad \text{with } L^{>0} \in \mathbb{R}^{k,c},\ L^0 \in \mathbb{R}^{k,r-c}, \\ D &= \begin{bmatrix} D^{>0} & 0_{c,r-c} \\ 0_{r-c,c} & 0_{r-c,r-c} \end{bmatrix}, \quad \text{with } D^{>0} \in \mathbb{R}^{c,c}, \\ R_2 &= \begin{bmatrix} R_2^{>0} & R_2^0 \end{bmatrix}, \quad \text{with } R_2^{>0} \in \mathbb{R}^{r,c},\ R_2^0 \in \mathbb{R}^{r,r-c}, \end{aligned}$$

so that :

$$Y^TXU = LDR_2^T = \begin{bmatrix} L^{>0} & L^0 \end{bmatrix} \begin{bmatrix} D^{>0} & 0_{c,r-c} \\ 0_{r-c,c} & 0_{r-c,r-c} \end{bmatrix} \begin{bmatrix} R_2^{>0,T} \\ R_2^{0,T} \end{bmatrix}.$$

Clearly, the columns of $L^{>0}$ are in $\text{Im } Y^TX$ since $D^{>0}R_2^{>0} \in \mathbb{R}^{c,r}$ is full-rank. Then we apply the Gram-Schmidt process to the columns of the matrix :

$$\begin{bmatrix} L^{>0} & M \end{bmatrix} \in \mathbb{R}^{k,c+r},$$

starting from the first column of M and until we obtain $r-c$ new orthogonal vectors. The computational cost is again $O(kr^2)$. Extracting these $r-c$ vectors and denoting $\bar{L} \in \mathbb{R}^{k,r-c}$ the matrix obtained by concatenation, we define $R_1 := \begin{bmatrix} L^{>0} & \bar{L} \end{bmatrix} \in \mathbb{R}^{k,r}$ and

$$Y^TXU = R_1DR_2^T = \begin{bmatrix} L^{>0} & \bar{L} \end{bmatrix} \begin{bmatrix} D^{>0} & 0_{c,r-c} \\ 0r-c,c & 0_{r-c,r-c} \end{bmatrix} \begin{bmatrix} R_2^{>0,T} \\ R_2^{0,T} \end{bmatrix}.$$

Thus we obtain a singular value decomposition $R_1DR_2^T$ of Y^TXU with $\text{Im } R_1 \subset \text{Im } Y^TX(X^TX)^{\frac{\dagger}{2}}$ at a computational cost of $O(kr^2)$. Eventually, given $U \in \mathbb{R}^{p,r}$, the subgradient of $U' \mapsto \left\|Y^TXU\right\|_*$ at U that we choose for Algorithm 1 is :

$$z_U = X^TYR_1R_2^T. \tag{G.6}$$

Remark 39. *In this paper, the condition Im $R_1 \subset$ Im $Y^TX(X^TX)^{\frac{\dagger}{2}}$ is only used in Lemma 40 to guarantee that $z_U \in \partial(-f_2)(U)$ where $f_2 : U' \mapsto \left\|Y^TXU'\right\|_*$. This property is then leveraged to prove global convergence for RRR and SRRR of the iterates produced by Algorithm 1 to a critical point in the sense of Definition 29. If we do not impose this extra condition and compute a subgradient as in Fact 37, all*

the results still hold except for the fact that we only guarantee global convergence to a KW-stationary point in the sense of Definition 30. When X^TX is invertible, we have shown that the induced computations have the same complexity $O(kr^2)$ as the computation of the SVD of Y^TXU.

Lemma 40. *Given $U \in \mathbb{R}^{p,r}$ let $R_1 D R_2^T$ be a singular value decomposition of Y^TXU with $R_1 \in \mathbb{R}^{k,r}$, $R_1^T R_1 = I_r$, Im $R_1 \subset$ Im $Y^TX(X^TX)^{\dagger\frac{1}{2}}$, $D \in \mathbb{R}^{r,r}$ a diagonal matrix with nonnegative coefficients, $R_2 \in \mathbb{R}^{r,r}$ and $R_2^T R_2 = I_r$. The matrix $-z_U := -X^TY R_1 R_2^T$ belongs to the limiting subdifferential presented in Definition 25 of the concave function $U' \mapsto -\left\|Y^TXU'\right\|_*$.*

Proof. First, with the notations of Lemma 40 and Proposition 6 of [Grave et al., 2011] that is recalled in Proposition 74, we know that when Y^TXU is full-rank, the function $f_2 : U' \mapsto \left\|Y^TXU'\right\|_*$ is differentiable at U with gradient $X^TY R_1 R_2^T$ so $-X^TY R_1 R_2^T \in \partial(-f_2)(U)$.

Secondly, we assume that Y^TXU has rank $c < r$. To prove that $-X^TY R_1 R_2^T \in \partial(-f_2)(U)$, we exhibit a sequence $(U_k)_{k\geq 0} \in (\mathbb{R}^{p,r})^{\mathbb{N}}$ such that, as in Definition 25,

$$U^k \to U,\ \left\|Y^tXU^k\right\|_* \to \left\|Y^TXU\right\|_*, \text{ and } X^TY R_1 R_2^T \in \hat{\partial}(-f_2)(U^k), \tag{G.7}$$

where $\hat{\partial}(-f_2)$ is the Fréchet subdifferential presented in Definition 25. Indeed, for $\epsilon > 0$, consider :

$$U_\epsilon := U + \epsilon (X^TX)^{\dagger\frac{1}{2}} P S^{-1} Q^T R_1 R_2^T,$$

where PSQ^T is the reduced singular value decomposition of $(X^TX)^{\dagger\frac{1}{2}} X^TY$. We have :

$$\begin{aligned} Y^TXU_\epsilon &= Y^TXU + \epsilon Y^TX(X^TX)^{\dagger\frac{1}{2}} P S^{-1} Q^T R_1 R_2^T \\ &= R_1 D R_2^T + \epsilon Q Q^T R_1 R_2^T \\ &= R_1 D R_2^T + \epsilon R_1 R_2^T \\ &= R_1 (D + \epsilon I_r) R_2^T. \end{aligned} \tag{G.8}$$

Equation (G.8) follows from $QQ^TR_1 = R_1$ because we assumed that Im $R_1 \subset$ Im $Y^TX(X^TX)^{\dagger\frac{1}{2}}$ and the columns of Q are an orthonormal basis of Im $Y^TX(X^TX)^{\dagger\frac{1}{2}}$. The trace norm is therefore differentiable at Y^TXU_ϵ that is full-rank and the gradient of $U' \mapsto \left\|Y^TXU'\right\|_*$ at U_ϵ is $X^TY R_1 R_2^T$. Defining $U_k := U_{\frac{1}{k}}$ for all $k > 0$ leads to (G.7). □

G.4.2 The proximal operator of the group-Lasso norm

In order to highlight the fact that U_+ is simply obtained by computing ∇f and the proximal operator of the group-Lasso norm, we could equivalently write Equation (G.3) as :

$$U_+ = \underset{U' \in \mathbb{R}^{p,r}}{\operatorname{argmin}} \frac{1}{2} \left\|U' - (U - t\nabla f(U))\right\|_F^2 + \lambda t \left\|U'\right\|_{1,2}. \tag{G.9}$$

An explicit form of this proximal operator is for instance given in Equation (3.7) in [Bach et al., 2012]. Given $1 \leq i \leq p$, let $[U_+]_{i,:}$ and $[U - t\nabla f(U)]_{i,:}$ denote the i-th

lines of the matrices U_+ and $U - t\nabla f(U)$ respectively. Assume that $[U - t\nabla f(U)]_{i,:} \neq 0$, then we have :

$$[U_+]_{i,:} = \max\left(0,\, 1 - \frac{\lambda t}{\|[U - t\nabla f(U)]_{i,:}\|_2}\right)[U - t\nabla f(U)]_{i,:}.$$

G.5 The Line Search Procedure in Algorithm 2

Given $t > 0$ and $U \in \mathbb{R}^{p,r}$, we recall the definitions of $\tilde{f}_{t,U}$, $\tilde{F}^{\lambda}_{t,U}$ and $\gamma_t(U)$:

$$\tilde{f}_{t,U}(U') = f(U) + \langle \nabla f(U), U' - U\rangle + \frac{1}{2t}\|U' - U\|_F^2\,,$$
$$\tilde{F}^{\lambda}_{t,U}(U') = \tilde{f}_{t,U}(U') + \lambda\|U'\|_{1,2}\,, \tag{G.10}$$
$$\gamma_t(U) = -\frac{1}{t}\min_{U'\in\mathbb{R}^d}\left[\tilde{F}^{\lambda}_{t,U}(U') - F^{\lambda}(U)\right]. \tag{G.11}$$

G.5.1 A lower-bound for the decrease in terms of function values

As announced in Section 5.5.3, we prove that $t\gamma_t(U)$ is a lower bound for the decrease at each iteration in terms of function values.

Fact 41. *Given $U \in \mathbb{R}^{p,r}$, t and U_+ obtained with Algorithm 2, the quantity $t\gamma_t(U)$ is a lower bound for the decrease in terms of function values from U to U_+ :*

$$t\gamma_t(U) \leq F^{\lambda}(U) - F^{\lambda}(U_+).$$

Proof. Indeed, we have :

$$t\gamma_t(U) = -\min_{U'\in\mathbb{R}^{p,r}}\left[\tilde{F}^{\lambda}_{t,U}(U') - F^{\lambda}(U)\right] \tag{G.12}$$
$$= F^{\lambda}(U) - \tilde{F}^{\lambda}_{t,U}(U_+) \tag{G.13}$$
$$\leq F^{\lambda}(U) - F^{\lambda}(U_+). \tag{G.14}$$

□

Equation (G.12) comes from the definition of γ_t in Equation (G.11). Equation (G.13) follows from the definition of U_+ in Equation (G.9). We have Equation (G.14) since the (LS) condition $\tilde{F}^{\lambda}_{t,U}(U^+) \geq F^{\lambda}(U^+)$ is satisfied for t and U_+.

G.5.2 A lower bound on the step size with the Line Search Procedure

In this section, we prove two additional results : that the (LS) condition is satisfied as soon as $t \leq \frac{1}{L_X}$ and, that there exists $\bar{k} \in \mathbb{N}$ such that for all $k \geq \bar{k}$, we have $t_k > \frac{\beta}{L_X}$.

Lemma 42. *Let $L_X > 0$ be the largest eigenvalue of X^TX. For any $t \leq \frac{1}{L_X}$ and $U, U' \in \mathbb{R}^{p,r}$, we have :*

$$f(U') + \lambda \|U'\|_{1,2} \leq f(U) + \langle \nabla f(U), U' - U \rangle + \frac{1}{2t} \|U' - U\|_F^2 + \lambda \|U'\|_{1,2} \quad \text{(G.15)}$$

where $y_U := X^TXU$ is the gradient of $U' \mapsto \frac{1}{2}\|XU'\|_F^2$, z_U is any subgradient of $U' \mapsto \left\|Y^TXU'\right\|_*$ and, with a slight abuse of notation, $\nabla f(U) := y_U - z_U$. Equivalently, for any $t \leq \frac{1}{L_X}$, the (LS) condition is satisfied i.e. we have

$$F^\lambda(U') \leq \tilde{F}^\lambda_{t,U}(U'). \quad \text{(G.16)}$$

In particular, Lemma 42 implies that Algorithm 2 terminates. This is illustrated in Figure G.1.

Proof. Let $U \in \mathbb{R}^{p,r}$. On the one hand, we have for all $U' \in \mathbb{R}^{p,r}$:

$$\begin{aligned}
\frac{1}{2}\|XU'\|_F^2 &= \frac{1}{2}\|X(U + (U' - U))\|_F^2 \\
&\leq \frac{1}{2}\|XU\|_F^2 + \langle X^TXU, U' - U \rangle + \frac{1}{2}\|X(U' - U)\|_F^2 \\
&\leq \frac{1}{2}\|XU\|_F^2 + \langle X^TXU, U' - U \rangle + \frac{L_X}{2}\|U' - U\|_F^2, \quad \text{(G.17)}
\end{aligned}$$

since $L_X > 0$ is the largest eigenvalue of X^TX. On the other hand, since z_U is a subgradient of $U' \mapsto \left\|Y^TXU'\right\|_*$, we have for any $U' \in \mathbb{R}^{p,r}$,

$$-\left\|Y^TXU'\right\|_* \leq -\left\|Y^TXU\right\|_* - \langle z_U, U' - U \rangle. \quad \text{(G.18)}$$

Summing Equation(G.17) and Equation(G.18), we obtain :

$$f(U') \leq \tilde{f}_{\frac{1}{L_X},U}(U').$$

Additionally, for any $0 < t \leq \frac{1}{L_X}$, we have :

$$\tilde{f}_{\frac{1}{L_X},U}(U') \leq \tilde{f}_{t,U}(U').$$

Consequently, for any $U, U' \in \mathbb{R}^{p,r}$ and $0 < t \leq \frac{1}{L_X}$, we have

$$F^\lambda(U') = f(U') + \lambda\|U'\|_{1,2} \leq \tilde{f}_{t,U}(U') + \lambda\|U'\|_{1,2} = \tilde{F}^\lambda_{t,U}(U'),$$

which is the (LS) condition.

□

Fact 43. *Let $k \geq 0$, $U_k \in \mathbb{R}^{p,r}$ and $t_{k-1} > 0$. Let $t > 0$ be defined as in Algorithm 2 : with probability $\pi \in (0, 1]$, t is set to $\frac{t_{k-1}}{\beta}$, otherwise, t is set to t_{k-1}. Let also $\bar{t}$ denote the initial step size, t_k and t_{k+1} the stepsizes produced by Algorithm 2 at iteration k and $k + 1$. We have the following properties :*

- *If $\frac{\beta}{L_X} < t \leq \frac{1}{L_X}$, then $t_k = t$ and $t_{k+1} = \frac{t}{\beta}$ or $t_{k+1} = t$ depending on the (LS) condition at iteration $k + 1$. In both cases, we have $t_{k+1} > \frac{\beta}{L_X}$.*

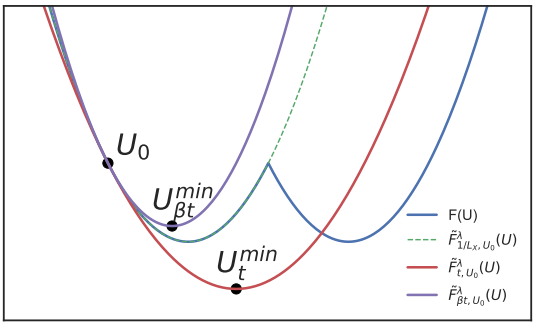


FIGURE G.1: Schematic representation of the Line Search Procedure in Algorithm 2. According to Equation (G.16), we have $\tilde{F}^{\lambda}_{1/L_X,U_0} \geq F^{\lambda}$, these two functions correspond to the dashed green line and the blue line. Given $U_0 \in \mathbb{R}^{p,r}$ and $t > 0$, we have represented $\tilde{F}^{\lambda}_{t,U_0}$ and $\tilde{F}^{\lambda}_{\beta t,U_0}$ under the assumptions $\beta t < \frac{1}{L_X} < t$ and $\tilde{F}^{\lambda}_{t,U}(U_t^{\min}) < F^{\lambda}(U_t^{\min})$ where $U_t^{\min}$ is the minimizer of $\tilde{F}^{\lambda}_{t,U_0}$. First, $U_t^{\min}$ is computed in Algorithm 2. As $\tilde{F}^{\lambda}_{t,U}(U_t^{\min}) < F^{\lambda}(U_t^{\min})$, the (LS) condition is not satisfied. The minimizer $U_{\beta t}^{\min}$ of $\tilde{F}^{\lambda}_{\beta t,U_0}$ is then computed and since, the (LS) condition is now satisfied, U_+ is set to $U_{\beta t}^{\min}$. Indeed, with Lemma 42, we are guaranteed to find $j \in \mathbb{N}$ such that $\tilde{F}^{\lambda}_{\beta^j t,U}(U_{\beta^j t}^{\min}) \geq F^{\lambda}(U_t^{\min})$ where $U_{\beta^j t}^{\min}$ is the minimizer of $\tilde{F}^{\lambda}_{\beta^j t,U}$.

- *If* $t \leq \frac{\beta}{L_X}$, *then* $t_k = t$ *and* $t_{k+1} = \frac{t}{\beta}$ *with probability* π, *otherwise* $t_{k+1} = t$.
- *If* $t > \frac{1}{L_X}$, *then* $t_k > \frac{\beta}{L_X}$.
- *For all* $k \geq 0$, *we have* $t_k \geq \min(\frac{\beta}{L_X}, \bar{t})$.

Proof. First, in Lemma 42, we have shown that the (LS) condition is satisfied as soon as $t \leq \frac{1}{L_X}$. Therefore, if $t \leq \frac{1}{L_X}$, the step is accepted in Algorithm 2 and $t_k = t$. At iteration $k+1$, the step size is set to $\frac{t_k}{\beta}$ with probability π and otherwise set to t_k. The step might only be rejected if the step size is set to $\frac{t_k}{\beta}$ and $\frac{t_k}{\beta} > \frac{1}{L_X}$. It would then be decreased by a multiplicative factor β and the step would be accepted with $t_{k+1} = t_k \times \frac{1}{\beta} \times \beta \leq \frac{1}{L_X}$.

Secondly, assume that $t \leq \frac{\beta}{L_X}$. Then $t \leq \frac{1}{L_X}$ since $\beta < 1$, the step is accepted in Algorithm 2 and $t_k = t$. At iteration $k+1$, t is set to $\frac{t_k}{\beta}$ with probability π and otherwise set to t_{k-1}. Anyway, we have at the next iteration $t \leq \frac{1}{L_X}$ so the (LS) condition is satisfied and the step is accepted.

Thirdly assume that $t > \frac{1}{L_X}$. By contradiction, suppose that $t_k \leq \frac{\beta}{L_X}$. The backtracking line search in Algorithm 2 ensures that there exists $j \in \mathbb{N}$ such that $t_k \leq \frac{\beta}{L_X} < \beta^j t \leq \frac{1}{L_X}$ and that the step size $\beta^j t$ was rejected because the (LS) condition was not satisfied. By Lemma 42, this is not possible since $\beta^j t \leq \frac{1}{L_X}$.

Consequently, if $t_k > \frac{\beta}{L_X}$ for a given $k \geq 0$, then for all $k' \geq k$ we have $t_{k'} > \frac{\beta}{L_X}$.

Thus, if $\bar{t} > \frac{\beta}{L_X}$ then for all $k \geq 0$, we have $t_k > \frac{\beta}{L_X}$. If $\bar{t} \leq \frac{\beta}{L_X}$, the algorithm progressively increases the value of t and after a few first iterations, say k, we have $t_k > \frac{\beta}{L_X}$: the step size t will be larger than $\frac{\beta}{L_X}$ after a number of steps which is finite in expectation. □

G.6 Study of the global convergence

Khamaru and Wainwright [2018] study the convergence of subgradient-type algorithms to KW-stationary points (see Definition 30) of non-convex and non-smooth functions that can be written as a sum of three terms $F = f_1 - f_2 + \lambda h$ where f_1 is a smooth function, f_2 is a continuous and convex function, h is a possibly non-smooth, convex penalty and $\lambda \geq 0$. Some of their results can be adapted to (RRR) and (SRRR) by taking :

$$\begin{aligned} f_1(U) &:= \frac{1}{2}\|XU\|_F^2\,, \\ f_2(U) &:= \left\|Y^T XU\right\|_*\,, \\ \text{and} \quad h(U) &:= \|U\|_{1,2}\,. \end{aligned}$$

First, we introduce the following results by Khamaru and Wainwright [2018] that we invoke in Section G.6.1 and Section G.6.2 .

Lemma 44. *[From Lemma 5 in Khamaru and Wainwright, 2018] Let $\lambda \geq 0$ and $(U_k)_{k\geq 0}$ be the sequence generated by Algorithm 1 and $(z_k)_{k\geq 0}$ the corresponding sequence of subgradients of f_2. For all $k \geq 0$, there is a subgradient s_{k+1} of $U \mapsto \|U\|_{1,2}$ at U_{k+1} such that :*

$$U_{k+1} = U_k - t_k\left[\nabla f_1(U_k) - z_k + \lambda s_{k+1}\right], \tag{G.19}$$

$$F^\lambda(U_k) - F^\lambda(U_{k+1}) \geq \frac{1}{2t_k}\|U_{k+1} - U_k\|_F^2\,. \tag{G.20}$$

Furthermore, for any convergent subsequence $(U_{k_j})_{j\geq 0}$ of the sequence $(U_k)_{k\geq 0}$ with $U_{k_j} \to \bar{U}$, we have :

$$\lim_{j\to+\infty} \left\|U_{k_j+1}\right\|_{1,2} = \left\|\bar{U}\right\|_{1,2}\,. \tag{G.21}$$

Lemma 44 is due to the choice of the forward-backward Algorithm 1 while the following Lemma 45 comes from the property of subanalytic functions [Attouch et al., 2010; Bolte et al., 2007, and references therein] given by Lemma 28. Indeed, norms and in particular the Frobenius norm, the trace-norm and the group-Lasso norm are subanalytic so the functions f and F^λ that we consider are subanalytic.

Lemma 45. *[From Lemma 6 in Khamaru and Wainwright, 2018] Let $\lambda \geq 0$, $(U_k)_{k\geq 0}$ be the sequence generated by Algorithm 1 and $(z_k)_{k\geq 0}$ the corresponding sequence of subgradients of f_2. The function F^λ is constant on the set of limit points $\bar{\mathcal{U}}$ of the sequence $(U_k)_{k\geq 0}$. We denote $\bar{F}^\lambda$ this limit. If we assume that $\bar{\mathcal{U}}$ contains only critical points of F^λ, then there exists constants $\theta \in [0,1)$, $C > 0$ and $k_1 \in \mathbb{N}$ such that for all $k \geq k_1$, we have :*

$$|F^\lambda(U_k) - \bar{F}^\lambda|^\theta \leq C\, dist(0, \nabla f_1(U_k) - z_{U_k} + \lambda\partial\|\cdot\|_{1,2}(U_k)). \tag{G.22}$$

G.6.1 Global convergence to a critical point with Algorithm 1 for RRR

The function $U \mapsto \frac{1}{2}\|XU\|_F^2$ is continuously differentiable and L_X-smooth where L_X is the largest eigenvalue of X^TX. The function $U \mapsto \left\|Y^TXU\right\|_*$ is continuous and convex and the difference $f(U) = \frac{1}{2}\|XU\|_F^2 - \left\|Y^TXU\right\|_*$ is bounded below by $-\frac{1}{2}\|Y\|_F^2$, indeed we have used in Section 5.3.1 the fact that for any $U \in \mathbb{R}^{p,r}$, we have :

$$\frac{1}{2}\|XU\|_F^2 - \left\|Y^TXU\right\|_* + \frac{1}{2}\|Y\|_F^2 = \min_{V\in\mathbb{R}^{k,r}:V^TV=I_r} \frac{1}{2}\left\|Y - XUV^T\right\|_F^2 \geq 0.$$

Besides, f satisfies the Kurdyka-Łojasiewicz property, presented in Definition 27, since it is the difference of two semi-algebraic functions. Therefore, our setting satisfies the conditions of Theorem 1 and Theorem 3 in Khamaru and Wainwright [2018] and we can prove that Algorithm 1 converges to a critical point from any initial point.

G.6.1.1 Limit points are critical points

The following result, whose proof is inspired from Theorem 1 by Khamaru and Wainwright [2018], ensures that any limit point $\bar{U}$ of the sequence generated by Algorithm 1 for RRR satisfies $0 \in \partial f(\bar{U})$.

Theorem 46. *Let $(U_k)_{k\geq 0}$ be the sequence generated by Algorithm 1 with $\lambda = 0$. The sequence of function values is decreasing and convergent. Besides, any limit point is a critical point of the function f.*

Proof. Equation (G.20) guarantees that the sequence of function values is decreasing. Since f has a finite lower-bound, the sequence of function values is convergent. Additionally, the iterates are bounded since the function is coercive *i.e.* $f(U) \to +\infty$ if $\|U\|_F \to \infty$.

To establish that the limit points are critical, consider a subsequence $(U_{k_j})_{j\geq 0}$ that converges to $\bar{U}$ and let $(z_{k_j})_{j\geq 0}$ be the associated subsequence of subgradients. Since the sequence $(U_{k_j})_{j\geq 0}$ converges to $\bar{U}$, we must have by Equation (G.19), $\left\|\nabla f_1(U_{k_j}) - z_{k_j}\right\|_F \to 0$. The function $f_1 : U \mapsto \frac{1}{2}\|XU\|_F^2$ being continuously differentiable, we have $\nabla f_1(U_{k_j}) \to \nabla f_1(\bar{U})$ and consequently $z_{k_j} \to \bar{z} := \nabla f_1(\bar{U})$. Besides, we know by Lemma 40 that for any $j \geq 0$, we have $-z_{k_j} \in \partial(-f_2)(U_{k_j})$.

We conclude like in the proof of Theorem 1 by Khamaru and Wainwright [2018], using the graph continuity of limiting subdifferentials which we recall in Fact 26, that $-\bar{z} \in \partial(-f_2)(\bar{U})$ and $\nabla f_1(\bar{U}) - \bar{z} = 0$, meaning that $0 \in \partial(f_1 - f_2)(\bar{U}) = \partial f(\bar{U})$. □

Remark 47. *Khamaru and Wainwright [2018] proved in an abstract but similar framework that the limit points are KW-stationary point in the sense of Definition 30, meaning that they can be stationary points for Algorithm 1. Instead, Theorem 46 guarantees, more standardly, that the limit points are critical in the sense that the limiting subdifferentials at these points contain the element 0. This is permitted by Lemma 40 which we obtained by imposing the condition Im $R_1 \subset$ Im Y^TX when computing a subgradient $X^TYR_1R_2^T$ of $U' \mapsto -\left\|Y^TXU'\right\|_*$, where $R_1DR_2^T$ is*

a singular value decomposition of Y^TXU. If the condition Im $R_1 \subset$ Im Y^TX was removed, exactly the same proof as for Theorem 46 would show that the limit points are KW-stationary points.

G.6.1.2 Convergence for RRR of Algorithm 1

Since f satifies the KŁ property, we can prove the convergence to a critical point.

Theorem 48. *[From Theorem 3 Khamaru and Wainwright, 2018] The sequence $(U_k)_{k\geq 0}$ produced by Algorithm 1 for RRR converges to a critical point.*

The proof of Theorem 48 is identical to the proof of Theorem 3 by Khamaru and Wainwright [2018]. We reproduce it here for completeness.

Proof. To prove that the sequence $(U_k)_{k\geq 0}$ has a finite length *i.e.* that we have

$$\sum_{k=0}^{+\infty} \|U_k - U_{k+1}\|_F < +\infty,$$

we use the KŁ property for subanalytic functions given by Lemma 45. Let $\theta \in [0, 1)$, $C > 0$, $k_1 \in \mathbb{N}$ be defined as in Lemma 45, $k \geq k_1$ and let $\bar{f}$ denote the limit of the sequence $\{f(U_k)\}_{k\geq 0}$. We have :

$$\begin{aligned}
(f(U_k) - \bar{f})^{1-\theta} - (f(U_{k+1}) - \bar{f})^{1-\theta} & \\
&\geq (1-\theta)(f(U_k) - \bar{f})^{-\theta} \left[f(U_k) - f(U_{k+1})\right] && \text{(G.23)} \\
&\geq \frac{(1-\theta)}{2t_k} (|f(U_k) - \bar{f}|)^{-\theta} \|U_k - U_{k+1}\|_F^2 && \text{(G.24)} \\
&\geq \frac{(1-\theta)}{2Ct_k \|\nabla f_1(U_k) - z_k\|_F} \|U_k - U_{k+1}\|_F^2 && \text{(G.25)} \\
&\geq \frac{(1-\theta)}{2C} \|U_k - U_{k+1}\|_F . && \text{(G.26)}
\end{aligned}$$

Inequality (G.23) follows from the concavity of $t \mapsto t^{1-\theta}$ and the inequalities $f(U_k) \geq f(U_{k+1}) \geq \bar{f}$. Inequality (G.24) comes from Equation (G.20) and the fact that $\{f(U_k)\}_{k\geq 0}$ is decreasing and converges to $\bar{f}$. Inequality (G.25) comes from Lemma 45. Finally, Inequality (G.26) follows from Equation (G.19). Summing both sides of Inequality (G.26) from $k = k_1$ to $k = +\infty$, we obtain :

$$\begin{aligned}
(f(U_{k_1}) - \bar{f})^{1-\theta} &= \sum_{k=k_1}^{+\infty} (f(U_k) - \bar{f})^{1-\theta} - (f(U_{k+1}) - \bar{f})^{1-\theta} \\
&\geq \frac{(1-\theta)}{2C} \sum_{k=k_1}^{+\infty} \|U_k - U_{k+1}\|_F ,
\end{aligned}$$

which proves the finite length property and the convergence of the sequence $(U_k)_{k\geq 0}$. With Theorem 46, we know that this limit is a critical point.

□

G.6.2 Global convergence to critical points with Algorithm 1 for SRRR

In this section, we justify global convergence of Algorithm 1 for SRRR to critical points and present conditions leveraged in Lemma 52 that ensure convergence to a unique point.

G.6.2.1 Limit points are critical points

The function f_1 is smooth and convex, the function f_2 is continuous and convex. In addition, the function $h : U \mapsto \|U\|_{1,2}$ is clearly proper, lower semi-continuous and convex and F^λ which is bounded below satisfies the KŁ property. Consequently, our setting for proximal gradient descent satisfies the conditions of the first part of Theorem 2 in Khamaru and Wainwright [2018] and we can adapt this result to SRRR.

Theorem 49. *Let $(U_k)_{k\geq 0}$ be the sequence generated by Algorithm 1 with $\lambda > 0$. The sequence of function values is decreasing and convergent. Besides, any limit point is a critical point of the function F^λ.*

Proof. Equation (G.20) guarantees that the sequence of function values is decreasing. Since F^λ has a finite lower-bound, the sequence of function values is convergent. Additionally, the iterates are bounded since the function is coercive *i.e.* $F^\lambda(U) \to +\infty$ if $\|U\|_F \to \infty$.

To establish that the limit points are critical, consider a subsequence $(U_{k_j})_{j\geq 0}$ that converges to $\bar{U} \in \mathbb{R}^{p,r}$ and let $(z_{k_j})_{j\geq 0}$ be the associated subsequence of subgradients, like in Equation (G.6). Since the sequence $(U_{k_j})_{j\geq 0}$ converges to $\bar{U}$ and f_2 is continuous, the sequence $\{f_2(U_k)\}_{k\geq 0}$ converges to $f_2(\bar{U})$. Given the form of the subgradients $(z_{k_j})_{j\geq 0}$ in Equation (G.6), they are bounded and we can assume, passing to a subsequence if necessary, that they converge to $\bar{z} \in \mathbb{R}^d$. Besides, we know by Lemma 40 that for any $j \geq 0$, we have $-z_{k_j} \in \partial(-f_2)(U_{k_j})$. Therefore, $\left\{(U_{k_j}, -z_{k_j}, -f_2(U_{k_j}))\right\}_{j\geq 0}$ converges to $(\bar{U}, -\bar{z}, -f_2(\bar{U}))$ and, using the graph continuity of limiting subdifferentials which we recall in Fact 26, we have $-\bar{z} \in \partial(-f_2)(\bar{U})$.

We now show that $-\nabla f_1(\bar{U}) + \bar{z} \in \partial(\lambda\|\cdot\|_{1,2})(\bar{U})$. Since $(\|U_{k_j} - U_{k_j+1}\|_F)_{j\geq 0}$ converges to zero, the sequence $(U_{k_j+1})_{j\geq 0}$ converges to $\bar{U}$ and by Equation (G.19), the sequence $(\|\nabla f_1(U_{k_j}) - z_{k_j} + \lambda s_{k_j+1}\|_F)_{j\geq 0}$ also converges to zero. Since f_1 is smooth, we know that $\left\{\nabla f_1(U_{k_j})\right\}_{j\geq 0}$ converges to $\nabla f_1(\bar{U})$. Combined with the convergence of $(z_{k_j})_{j\geq 0}$ to $\bar{z}$, it shows that $(\lambda s_{k_j+1})_{j\geq 0}$ converges to $\lambda\bar{s} := -\nabla f_1(\bar{U}) + \bar{z}$. With Equation (G.21) in Lemma 44, we also have that $(\lambda\|U_{k_j+1}\|_{1,2})_{j\geq 0}$ converges to $\lambda\|\bar{U}\|_{1,2}$. All this leads to the convergence of $\left\{(U_{k_j+1}, \lambda s_{k_j+1}, \lambda\|U_{k_j+1}\|_{1,2})\right\}_{j\geq 0}$ to $(\bar{U}, \lambda\bar{s}, \lambda\|\bar{U}\|_{1,2})$. Consequently, the graph continuity in Fact 26 guarantees that $\lambda\bar{s} \in \partial(\lambda\|\cdot\|_{1,2})(\bar{U})$. Finally, we conclude that $\nabla f_1(\bar{U}) - \bar{z} + \lambda\bar{s} = 0 \in \partial F^\lambda(\bar{U})$ *i.e.* $\bar{U}$ is a critical point of F^λ. □

Remark 50. *The same comments as in Remark 47 hold for Theorem 49 and the comparison between its proof and the proof of Theorem 2 by Khamaru and Wainwright [2018].*

G.6.2.2 Convergence for SRRR of Algorithm 1

In order to prove convergence of the sequence $(U_k)_{k\geq 0}$, Theorem 4 of Khamaru and Wainwright [2018] formally requires that f_2 is a smooth function, a requirement which is not met by $U \mapsto \left\|Y^T X U\right\|_*$. Nonetheless, an inspection of the proof shows that local smoothness in a neighborhood of the critical points of the function is sufficient. More precisely, the same proof as for Theorem 4 in Khamaru and Wainwright [2018] can be used for SRRR as long as we can guarantee that there exists $k_1 \geq 0$ such that for all $k \geq k_1$, the iterates $(U_k)_{k\geq k_1}$ lie in a compact subset where f is locally smooth. We denote $\bar{\mathcal{U}}$ the set of limit points of the sequence $(U_k)_{k\geq 0}$ and for any $S \subset \{1, \ldots, p\}$, X_S is the matrix formed by keeping the columns of X indexed by S.

Theorem 51. *Assume that*

- $\mathcal{H}1$ *: The step sizes* $(t_k)_{k\geq 0}$ *produced by Algorithm 1 are upper bounded by a constant* $d > 0$.
- $\mathcal{H}2$ *: The set of limit points* $\bar{\mathcal{U}}$ *of the sequence produced by Algorithm 1 is a subset of the local minima of* F^λ *and contains only matrices with at least* r *non-zero rows.*
- $\mathcal{H}3$ *: For any* $S \subset \{1, \ldots, p\}$ *of cardinality at least* r*, the matrix* $X_S^T Y$ *is full-rank.*

Then the sequence $(U_k)_{k\geq 0}$ *produced by Algorithm 1 for SRRR converges to a critical point.*

The assumptions $\mathcal{H}1$ and $\mathcal{H}2$ are used in the proof of Theorem 51. The assumption $\mathcal{H}2$ will hold unless local minima are so sparse that the number of selected variables is strictly smaller than r in which case the rank constraint becomes essentially useless. The assumption $\mathcal{H}3$ will hold with probability one if X and Y contain for example additive noise. It is leveraged in Appendix G.11.1 to prove Lemma 52 that we introduce below with Lemma 53 and Lemma 54 before giving the proof of Theorem 51.

Lemma 52. *With Assumption* $\mathcal{H}3$*, any local minimum* U^* *of (SRRR) which has at least* r *non-zero rows must be full-rank.*

Put differently, Assumption $\mathcal{H}2$ and Assumption $\mathcal{H}3$ combined with Lemma 52 imply that the set of limit points $\bar{\mathcal{U}}$ contains only full-rank matrices. The next lemma ensures that the function $f_2 : U \mapsto \left\|Y^T X U\right\|_*$ is differentiable at such points, it is proved in Appendix G.11.2.

Lemma 53. *With Assumption* $\mathcal{H}3$*, let* U^* *be a full-rank local minimum of (SRRR). Necessarily,* $Y^T X U^*$ *is full-rank.*

Lemma 53 is essential to prove locally a Lipschitz gradients property which is formalized in Lemma 54, proved in Appendix G.11.3.

Lemma 54. *With Assumption* $\mathcal{H}2$ *and Assumption* $\mathcal{H}3$*, there exists* $M > 0$ *and* $k_1 \geq 0$ *such that for any* $k \geq k_1$*,* f *is differentiable at* U_k*,* U_{k+1} *and we have :*

$$\left\|\nabla f(U_k) - \nabla f(U_{k+1})\right\|_F \leq M \left\|U_k - U_{k+1}\right\|_F . \tag{G.27}$$

Proof of Theorem 51 . Let $k_1 \geq 0$ be defined as in Lemma 54. For $k \geq k_1$ we denote z_k a gradient of f_2 obtained through the update in Algorithm 1 and s_k a subgradient of $U \mapsto \lambda \|U\|_{1,2}$ at U_k. Let $k > k_1$, we have :

$$\begin{aligned}
&\|\nabla f_1(U_k) - z_k + \lambda s_k\|_F \\
&= \left\| (\nabla f_1(U_k) - z_k) + (z_{k-1} - \nabla f_1(U_{k-1})) + \frac{1}{t_{k-1}}(U_{k-1} - U_k) \right\|_F && \text{(G.28)} \\
&\leq \|(\nabla f_1(U_k) - z_k) - (\nabla f_1(U_{k-1}) - z_{k-1})\|_F + \frac{1}{t_{k-1}} \|U_{k-1} - U_k\|_F && \text{(G.29)} \\
&= \|\nabla f(U_k) - \nabla f(U_{k-1})\|_F + \frac{1}{t_{k-1}} \|U_{k-1} - U_k\|_F \\
&\leq (M + \frac{1}{t_{k-1}}) \|U_k - U_{k-1}\|_F . && \text{(G.30)}
\end{aligned}$$

Inequality (G.28) follows from the update in Algorithm 1. Inequality (G.29) comes from the triangle inequality. Inequality (G.30) is due to Equation (G.27).

The second argument we give is similar to Equation (G.26) in the proof of Theorem 48. Since the functions we consider are subanalytic, we can consider $\theta \in [0, 1)$, $C > 0$ and $k_2 \geq k_1$ defined as in Lemma 45. Let $\bar{F}^\lambda$ denote the limit of the sequence $\left\{F^\lambda(U_k)\right\}_{k\geq 0}$ and $k \geq k_2$, we have :

$$\begin{aligned}
&(F^\lambda(U_k) - \bar{F}^\lambda)^{1-\theta} - (F^\lambda(U_{k+1}) - \bar{F}^\lambda)^{1-\theta} \\
&\qquad \geq (1-\theta) \left[F^\lambda(U_k) - \bar{F}^\lambda\right]^{-\theta} \left[F^\lambda(U_k) - F^\lambda(U_{k+1})\right] && \text{(G.31)} \\
&\qquad \geq \frac{(1-\theta)}{2t_k} \left[|F^\lambda(U_k) - \bar{F}^\lambda|\right]^{-\theta} \|U_k - U_{k+1}\|_F^2 && \text{(G.32)} \\
&\qquad \geq \frac{(1-\theta)}{2Ct_k \|\nabla f(U_k) + \lambda s_k\|_F} \|U_k - U_{k+1}\|_F^2 && \text{(G.33)} \\
&\qquad \geq \frac{(1-\theta)}{2Cd \|\nabla f(U_k) + \lambda s_k\|_F} \|U_k - U_{k+1}\|_F^2 . && \text{(G.34)}
\end{aligned}$$

Inequality (G.31) follows from the concavity of $t \mapsto t^{1-\theta}$ and the inequalities $F^\lambda(U_k) \geq F^\lambda(U_{k+1}) \geq \bar{F}^\lambda$. Inequality (G.32) comes from Equation (G.20) and the fact that $\left\{F^\lambda(U_k)\right\}_{k\geq 0}$ is decreasing and converges to $\bar{F}^\lambda$. Inequality (G.33) comes from Lemma 45. Finally, Inequality (G.34) follows from Assumption $\mathcal{H}1$ in Theorem 51. Combining Inequality (G.30) with Inequality (G.34), we obtain :

$$\begin{aligned}
&(F^\lambda(U_k) - \bar{F}^\lambda)^{1-\theta} - (F^\lambda(U_{k+1}) - \bar{F}^\lambda)^{1-\theta} \\
&\qquad \geq \frac{(1-\theta)}{2Cd(M + \frac{1}{t_{k-1}})} \frac{\|U_k - U_{k+1}\|_F^2}{\|U_{k-1} - U_k\|_F} \\
&\qquad \geq \frac{(1-\theta)}{2Cd(M + \frac{1}{\min(\frac{\beta}{L_X}, t_{-1})})} \frac{\|U_k - U_{k+1}\|_F^2}{\|U_{k-1} - U_k\|_F} . && \text{(G.35)}
\end{aligned}$$

Equation (G.35) follows from Fact 43. The rest of the proof leads to the finite length property and completely follows the proof of Theorem 4 in Khamaru

and Wainwright [2018] since they also leverage only the local property of Lipschitz gradients in a compact set. We denote :

$$\Delta_k := C' \left[(F^\lambda(U_k) - \bar{F}^\lambda)^{1-\theta} - (F^\lambda(U_{k+1}) - \bar{F}^\lambda)^{1-\theta} \right], \tag{G.36}$$

$$\text{where} \quad C' := \frac{2Cd\left[M + \max(\frac{L_X}{\beta}, \frac{1}{t_{-1}})\right]}{(1-\theta)}.$$

Equation (G.35) can be rewritten :

$$\|U_k - U_{k+1}\|_F \leq \sqrt{\Delta_k \|U_{k-1} - U_k\|_F}.$$

Summing from $j = k_2 + 1$ to $j = k$, we obtain :

$$\begin{aligned}
\sum_{j=k_2+1}^{k} \|U_j - U_{j+1}\|_F &\leq \sum_{j=k_2+1}^{k} \sqrt{\Delta_j \|U_{j-1} - U_j\|_F} \\
&\leq \sum_{j=k_2+1}^{k} \frac{1}{2}\Delta_j + \frac{1}{2}\|U_{j-1} - U_j\|_F \qquad \text{(G.37)} \\
&\leq \frac{C'}{2}(F^\lambda(U_{k_2+1}) - \bar{F}^\lambda)^{1-\theta} + \frac{1}{2}\sum_{j=k_2+1}^{k} \|U_{j-1} - U_j\|_F . \qquad \text{(G.38)}
\end{aligned}$$

Inequality (G.37) follows from the inequality of arithmetic and geometric means. Inequality (G.38) comes from Equation (G.36). Rewriting Inequality (G.38), we have for all $k \geq k_2 + 2$,

$$\begin{aligned}
&\left[\frac{1}{2}\sum_{j=k_2+1}^{k-1} \|U_j - U_{j+1}\|_F\right] + \left[\frac{1}{2}\sum_{j=k_2+2}^{k} \|U_{j-1} - U_j\|_F\right] + \|U_k - U_{k+1}\|_F \\
&\leq \frac{C'}{2}(F^\lambda(U_{k_2+1}) - \bar{F}^\lambda)^{1-\theta} + \left[\frac{1}{2}\sum_{j=k_2+2}^{k} \|U_{j-1} - U_j\|_F\right] + \frac{1}{2}\|U_{k_2} - U_{k_2+1}\|_F .
\end{aligned}$$

This last inequality implies that :

$$\begin{aligned}
&\frac{1}{2}\sum_{j=k_2+1}^{k-1} \|U_j - U_{j+1}\|_F \\
&\qquad \leq \frac{C'}{2}(F^\lambda(U_{k_2+1}) - \bar{F}^\lambda)^{1-\theta} + \frac{1}{2}\|U_{k_2} - U_{k_2+1}\|_F - \|U_k - U_{k+1}\|_F \\
&\qquad \leq \frac{C'}{2}(F^\lambda(U_{k_2+1}) - \bar{F}^\lambda)^{1-\theta} + \frac{1}{2}\|U_{k_2} - U_{k_2+1}\|_F \\
&\qquad < +\infty.
\end{aligned}$$

Eventually, we conclude that the sequence $(U_k)_{k\geq 0}$ has finite length and therefore converges to an element $\bar{U} \in \mathbb{R}^{p,r}$. With Theorem 49, we know that $\bar{U}$ is a critical point. □

G.7 Proofs for section 5.5.1

In this section, we are going to prove Equation (5.7), Lemma 9 and Lemma 10. We maintain the following assumptions :

$$r \leq \ell, \tag{G.39}$$

$$s_1 > \ldots > s_\ell. \tag{G.40}$$

At first, to widen the scope of our results, we will not make the assumption :

$$X^T X \text{ is invertible.} \tag{G.41}$$

Assumption (G.41) will play a key role in the analysis and impact the results. We will precise what it implies for the analysis when it is satisfied and when it is not.

G.7.1 Proof of Equation (5.7)

While we assumed that X is full-rank in the core of the article, we do not make this assumption in this section to prove a more general result than Equation (5.7). Of course, the latter can be obtained as a special case. Let $m \leq p$ be the rank of X and consider :

$$KD^2K^T \text{ the reduced singular value decomposition of } X^TX,$$

with $K \in \mathbb{R}^{p,m}$, $K^TK = I_m$ and $D \in \mathbb{R}^{m,m}$ a diagonal matrix with positive entries on the diagonal. We also write :

$$(X^TX)^\dagger := KD^{-2}K^T \text{ the pseudo-inverse of } X^TX,$$

$$(X^TX)^{\frac{\dagger}{2}} := KD^{-1}K^T \text{ a square-root of } (X^TX)^\dagger,$$

and :

$$(X^TX)^{\frac{1}{2}} := KDK^T \text{ a square root of } X^TX.$$

Let :

$$K^\perp \in \mathbb{R}^{p,p-m} \text{ such that } \begin{bmatrix} K & K^\perp \end{bmatrix}^T \begin{bmatrix} K & K^\perp \end{bmatrix} = I_p.$$

Here, we denote PSQ^T the reduced singular values of $(X^TX)^{\frac{\dagger}{2}}X^TY$, with $\ell := \text{rank}(X^TX)^{\frac{\dagger}{2}}X^TY \leq \min(m,k)$, $P \in \mathbb{R}^{p,\ell}$, $S \in \mathbb{R}^{\ell,\ell}$ and $Q \in \mathbb{R}^{\ell,k}$. We also define $P^\perp \in \mathbb{R}^{p,m-\ell}$ such that the columns of the matrix $\begin{bmatrix} P & P^\perp \end{bmatrix}$ form an orthonormal basis of Im X^T. If X is full-rank, this definition corresponds indeed with the matrices that were introduced in Section 5.5.1. The definition of τ is in this more general case :

$$\tau : \begin{cases} \mathbb{R}^{\ell,r} \times \mathbb{R}^{m-\ell,r} \times \mathbb{R}^{p-m,r} \to \mathbb{R}^{p,r} \\ (A,C,N) \mapsto (X^TX)^{\frac{\dagger}{2}} \begin{bmatrix} P & P^\perp \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & I_{m-\ell} \end{bmatrix} \begin{bmatrix} A \\ C \end{bmatrix} + K^\perp N \end{cases}. \tag{G.42}$$

Of course, under the additional assumption that X^TX is invertible, the term $K^\perp N$ would be removed and τ would be the same as the one we defined in Equation (5.6).

We define $f_{a,c,n} := f \circ \tau$ and we prove that :

$$f_{a,c,n}(A, C, N) = \frac{1}{2}\|SA\|_F^2 - \|S^2 A\|_* + \frac{1}{2}\|C\|_F^2. \tag{G.43}$$

Equation (5.7) can be obtained similarly if $X^T X$ is invertible.

Proof of Equation (G.43). Let $(A, C, N) \in \mathbb{R}^{\ell,r} \times \mathbb{R}^{m-\ell,r} \times \mathbb{R}^{p-m,r}$, we have :

$$\begin{aligned}
f_{a,c,n}(A, C, N) &= f \circ \tau(A, C, N) && \text{(G.44)}\\
&= f((X^T X)^{\frac{\dagger}{2}}(PSA + P^\perp C) + K^\perp N) && \text{(G.45)}\\
&= \frac{1}{2}\|(X^T X)^{\frac{1}{2}}((X^T X)^{\frac{\dagger}{2}}(PSA + P^\perp C) + K^\perp N)\|_F^2 \\
&\quad - \|Y^T X((X^T X)^{\frac{\dagger}{2}}(PSA + P^\perp C) + K^\perp N)\|_* && \text{(G.46)}\\
&= \frac{1}{2}\|(X^T X)^{\frac{1}{2}}(X^T X)^{\frac{\dagger}{2}}(PSA + P^\perp C)\|_F^2 \\
&\quad - \|Y^T X(X^T X)^{\frac{\dagger}{2}}(PSA + P^\perp C)\|_* && \text{(G.47)}\\
&= \frac{1}{2}\|PSA\|_F^2 + \frac{1}{2}\|P^\perp C\|_F^2 - \|QSP^T(PSA + P^\perp C)\|_*, && \text{(G.48)}\\
&= \frac{1}{2}\|SA\|_F^2 + \frac{1}{2}\|C\|_F^2 - \|QS^2 A\|_* && \text{(G.49)}\\
&= \frac{1}{2}\|SA\|_F^2 - \|S^2 A\|_* + \frac{1}{2}\|C\|_F^2. && \text{(G.50)}
\end{aligned}$$

□

Equation (G.44) follows from the definition of $f_{a,c,n}$ and Equation (G.45) from the definition of τ. Equation (G.46) follows from the definition of f and since for all $M \in \mathbb{R}^{p,r}$, we have $\|XM\|_F^2 = \|(X^T X)^{\frac{1}{2}} M\|_F^2$. We have Equation (G.47) since $XK^\perp = 0$. Equation (G.48) comes from the facts that P, $P^\perp \in \operatorname{Im} X$ and $(X^T X)^{\frac{1}{2}}(X^T X)^{\frac{\dagger}{2}}$ acts like the identity on $\operatorname{Im} X^T$ for the first term and $QSP^T = Y^T X(X^T X)^{\frac{\dagger}{2}}$ for the second term. We have Equation (G.49) because $\begin{bmatrix} P & P^\perp \end{bmatrix}$ is orthogonal and Equation (G.50) because the columns of Q are orthogonal.

G.7.2 Proof of Lemma 9

We denote Ω_a^* the set of minima of $f_a : A \in \mathbb{R}^{\ell,r} \mapsto \frac{1}{2}\|SA\|_F^2 - \|S^2 A\|_*$ where $S = \operatorname{diag}(s_1 > \ldots > s_\ell) \in \mathbb{R}^{\ell,\ell}$. To prove that $\Omega_a^* = \left\{\tilde{I}R \mid R \in \mathcal{O}_r\right\}$ with $\tilde{I} = (1_{i=j})_{1 \le i \le \ell,\, 1 \le j \le r} \in \mathbb{R}^{\ell,r}$, first note that the two following problems have the same optimal value :

$$\min_{A \in \mathbb{R}^{\ell,r},\ V \in \mathbb{R}^{\ell,r}} f_{a,v}(A, V) \text{ where } f_{a,v}(A, V) := \frac{1}{2}\|S - SAV^T\|_F^2, \tag{G.51}$$

$$\min_{A \in \mathbb{R}^{\ell,r},\ V \in \mathbb{R}^{\ell,r} :\, V^T V = I_r} f_{a,v}(A, V). \tag{G.52}$$

Indeed, for any A, $V \in \mathbb{R}^{\ell,r}$, there exists A', $V' \in \mathbb{R}^{\ell,r}$ such that $V^T V = I_r$ and $AV^T = A'V'^T$. For instance, the matrices can be obtained from the singular value

decomposition $R_1 \Sigma R_2^T$ of AV^T by taking $A' = R_1\Sigma$ and $V' = R_2$. Besides, given $A \in \mathbb{R}^{\ell,r}$ and $V \in \mathbb{R}^{\ell,r}$, we have :

$$f_{a,v}(A,V) = \frac{1}{2}\|S - SAV^T\|_F^2 = \frac{1}{2}\|S\|_F^2 + \frac{1}{2}\|SAV^T\|_F^2 - \langle S, SAV^T\rangle.$$

Defining $V_A \in \operatorname{argmax}_{V\in\mathbb{R}^{\ell,r}:\, V^TV=I_r} \langle S, SAV^T\rangle$ and using Fact 33, we obtain :

$$\frac{1}{2}\|S - SAV_A^T\|_F^2 = \frac{1}{2}\|S\|_F^2 + \frac{1}{2}\|SA\|_F^2 - \left\|S^2A\right\|_*.$$

Consequently, if A is a minimizer of :

$$\min_{A\in\mathbb{R}^{\ell,r}} f_a(A), \tag{G.53}$$

where $f_a(A) = \frac{1}{2}\|SA\|_F^2 - \|S^2A\|_*$, then (A, V_A) is a minimizer of Problem (G.51). This means in particular that SAV_A^T is a minimizer of :

$$\min_{M\in\mathbb{R}^{\ell,\ell}:\,\operatorname{rank}(M)\le r} \frac{1}{2}\|S - M\|_F^2.$$

The matrix SAV_A^T must be equal to the best low-rank approximation for the Frobenius norm of S and, by the Eckart-Young-Mirsky theorem, this best approximation is $S\tilde{I}\tilde{I}^T$ with $\tilde{I} = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \in \mathbb{R}^{\ell,r}$ since we have assumed that the values on the diagonal of S are strictly decreasing.

The matrix S is invertible so we must have $AV^T = \tilde{I}\tilde{I}^T$ which is equivalent, if $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ and $V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$ with A_1, $V_1 \in \mathbb{R}^{r,r}$ and A_2, $V_2 \in \mathbb{R}^{\ell-r,\ell-r}$, to :

$$\begin{bmatrix} A_1V_1^T & A_1V_2^T \\ A_2V_1^T & A_2V_2^T \end{bmatrix} = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}. \tag{G.54}$$

The second line $A_2V^T = 0$ implies $A_2 = 0$ since $V^TV = I_r$. From the first line of the matrices in Equation (G.54), $A_1V_1^T = I_r$ implies that A_1 is invertible so $A_1V_2^T = 0$ implies that $V_2 = 0$ and A_1 has to be orthogonal as it is the inverse of V_1^T. Put differently, $A^T = \begin{bmatrix} V_1^T & 0_{r,\ell-r} \end{bmatrix} = V_1^T\tilde{I}^T$ where V_1 is an orthogonal square matrix *i.e* an orthogonal matrix. Thus, any optimum A belongs to $\Omega_a^* := \left\{\tilde{I}R \mid R \in \mathcal{O}_r\right\}$. Conversely, for any $R \in \mathcal{O}_r$ we have $f_a(\tilde{I}R) = \frac{1}{2}\left\|S\tilde{I}\right\|_F^2 - \left\|S^2\tilde{I}\right\|_*$: this implies that all the elements in Ω_a^* are optima.

G.7.3 Proof of Lemma 10

We show that all local minima of f_a are global. The result is the same for f given that $f \circ \tau(A,C) = f_a(A) + \frac{1}{2}\|C\|_F^2$ and τ is the invertible linear transformation defined in Equation (G.42). First we start by eliminating the possibility of having a local maximum other than 0 with the following result.

Lemma 55. *Only* 0 *can be a local maximum of* f_a.

Proof. For any A, the restriction of f_a to the one-dimensional set $\mathcal{D}_A := \{\alpha A,\, \alpha \geq 0\}$ is a convex polynomial function of degree 2. Indeed, for any $\alpha > 0$, we have

$$f_a(\alpha A) = \frac{\alpha^2}{2}\|SA\|_F^2 - \alpha\|S^2 A\|_*.$$

Since $S \in \mathbb{R}^{\ell,\ell}$ is an invertible diagonal matrix, only 0 can be a local maximum of f_a. □

Corollary 56. *The zero matrix is indeed a local maximum of the function f_a.*

Proof. Thanks to the equivalence of norms in finite dimensions and the fact that S has only positive elements on its diagonal, we know that there exists $c, d > 0$ such that for any $A \in \mathbb{R}^{\ell,r}$, $t > 0$, we have :

$$f_a(0 + tA) \leq c\,\|A\|_F^2\, t^2 - d\,\|A\|_F\, t.$$

The zero matrix is necessary a local maximum. □

To deal with critical points, we treat separately rank-deficient matrices and full-rank matrices. The following result, proved in Appendix G.11.4, considers the case of rank-deficient matrices.

Lemma 57. *Let $A \in \mathbb{R}^{\ell,r}$ be a rank-deficient matrix, there exists $B \in \mathbb{R}^{\ell,r}$ such that $\|B\|_F = 1$ and $\delta > 0$ such that for all $-\delta < t < \delta$, we have :*

$$f_a(A + tB) \leq f_a(A) - \frac{s_\ell^2}{2}|t|.$$

Therefore, no rank-deficient matrix can be a local minimum of f_a.

In order to deal with full-rank matrices and having already described the set of optima, we characterize the set of full-rank critical points. Consider the set $\mathcal{P}$ of permutations $\pi : [\![1; \ell]\!] \to [\![1; \ell]\!]$ such that $\pi(1) < \ldots < \pi(r)$ and simultaneously $\pi(r+1) < \ldots < \pi(\ell)$. For any $\pi \in \mathcal{P}$, we denote :

$$\Pi_\pi := (1_{i=\pi(j)})_{1\leq i\leq \ell,\, 1\leq j\leq r} \in \mathbb{R}^{\ell,r}. \tag{G.55}$$

Note that the sole purpose of the condition $\pi(r+1) < \cdots < \pi(\ell)$ is to have a one-to-one correspondence between the set of permutations $\mathcal{P}$ and the set of matrices $\{\Pi_\pi \mid \pi \in \mathcal{P}\}$. We have the following result, proved in Appendix G.11.5.

Lemma 58. *If the values of the diagonal matrix S are strictly decreasing i.e. $S = diag\,(s_1 > \ldots > s_\ell)$, then the set Ω_a^s of differentiable critical points for problem (G.53) is the image by linear transformations from $\mathbb{R}^{r,r}$ to $\mathbb{R}^{\ell,r}$ of $\mathcal{O}_r$:*

$$\Omega_a^s = \{\Pi_\pi R |\ \pi \in \mathcal{P}, R \in \mathcal{O}_r\}.$$

Besides, Ω_a^s contains only global minima and saddle points.

We could have an even more precise description of the behavior of f_a around the saddle points with Theorem 63 and Corollary 65 (given below). Saddle points are in fact strict saddle points *i.e.* the Hessian at these points has at least one negative eigenvalue. However, that is not necessary here.

We can now prove Lemma 10.

Proof of Lemma 10. We know from Lemma 57 that a rank-deficient matrix cannot be a local minimum. The function f_a is differentiable at $A \in \mathbb{R}^{\ell,r}$ if and only if A is full-rank[1]. Finally, Lemma 58 details all critical points where A is full-rank, they are either global minima or saddle points. □

G.8 Proofs for Section 5.5.2

G.8.1 Proof of Lemma 11

In Section 5.5.2, we have introduced for any $A \in \mathbb{R}^{p,r}$,

$$\Pi_{\Omega_a^*}(A) := \operatorname*{argmin}_{B \in \Omega_a^*} \|B - A\|_F^2,$$

and :

$$\mathcal{C}_a(R) := \{A \in \mathbb{R}^{\ell,r} \mid \tilde{I}R \in \Pi_{\Omega_a^*}(A)\}. \tag{G.56}$$

First, we prove Equation (5.8) that describes $\mathcal{C}_a(I_r)$. According to Lemma 9, $\Omega_a^* = \left\{\tilde{I}R \mid R \in \mathcal{O}_r\right\}$, so with Fact 32, we could have equivalently defined $\Pi_{\Omega_a^*}(A)$ with $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$, $A_1 \in \mathbb{R}^{r,r}$ and $A_2 \in \mathbb{R}^{\ell-r,r}$ as :

$$\operatorname*{argmin}_{\tilde{I}R:\, R \in \mathcal{O}_r} \|\tilde{I}R - A\|_F^2 = \operatorname*{argmax}_{\tilde{I}R:\, R \in \mathcal{O}_r} \langle \tilde{I}R, A\rangle = \tilde{I} \operatorname*{argmax}_{R \in \mathcal{O}_r} \langle R, \tilde{I}^T A\rangle = \tilde{I} \operatorname*{argmax}_{R \in \mathcal{O}_r} \langle R, A_1\rangle. \tag{G.57}$$

By Fact 36, we have that $I_r \in \operatorname{argmax}_{R:\, R \in \mathcal{O}_r} \langle R, A_1\rangle$ if and only if A_1 is positive-semidefinite. This proves Equation (5.8).

Secondly, the equality $\mathcal{C}_a(R) = \{AR \mid A \in \mathcal{C}_a(I_r)\}$ basically stems from the definition of $\Pi_{\Omega_a^*}$ since :

$$\|\tilde{I} - A\|_F^2 = \|\tilde{I}R - AR\|_F^2. \tag{G.58}$$

Indeed, Equation (G.58) implies that $A \in \mathcal{C}_a(I_r)$ if and only if $AR \in \mathcal{C}_a(R)$.

Finally, to prove that $\cup_{R \in \mathcal{O}_r} \mathcal{C}_a(R) = \mathbb{R}^{\ell,r}$, consider $M \in \mathbb{R}^{\ell,r}$ and :

$$B_M \in \operatorname*{argmin}_{B \in \Omega_a^*} \|B - M\|_F^2.$$

According to Lemma 9, $\Omega_a^* := \left\{\tilde{I}R \mid R \in \mathcal{O}_r\right\}$ is compact. Therefore, there exists $R \in \mathcal{O}_r$ such that $B_M = \tilde{I}R$. Obviously, the definition of $\mathcal{C}_a(R)$ given in Equation (G.56) implies that $M \in \mathcal{C}_a(R)$.

The following fact gives more details on the structure of the cones that we built.

Fact 59. *The relative interiors[2] of all the cones partition the set of matrices* $[A_1^T \ A_2^T]^T$ *such that* $A_1 \in \mathbb{R}^{r,r}$ *is invertible and* $A_2 \in \mathbb{R}^{\ell-r,r}$.

[1] Details about the derivative of the trace-norm are given in Proposition 74.

[2] Given a set in a Euclidean space, its relative interior is the interior of this set within the subspace spanned by its elements.

Proof. First, since the relative interior of the set $\mathcal{S}_r^+$ of positive-semidefinite matrices is the set $\mathcal{S}_r^{++}$ of positive-definite matrices, given $R \in \mathcal{O}_r$, the relative interior of the cone $\mathcal{C}_a(R)$ is the set :

$$\left\{ \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} R \mid A_1 \in \mathcal{S}_r^{++},\ A_2 \in \mathbb{R}^{\ell-r,r} \right\}.$$

Secondly, according to Equation (G.57), the matrix $A = [A_1^T\ A_2^T]^T \in \mathcal{C}_a(R)$ with $R \in \mathcal{O}_r$ if and only if $R \in \operatorname{argmax}_{R' \in \mathcal{O}_r} \langle R', A_1 \rangle$. According to Fact 35, there is a unique element in $\operatorname{argmax}_{R' \in \mathcal{O}_r} \langle R', A_1 \rangle$ if A_1 has full rank. Therefore, given $[A_1^T\ A_2^T]^T$ such that $A_1 \in \mathbb{R}^{r,r}$ is invertible and $A_2 \in \mathbb{R}^{\ell-r,r}$, there is a unique $R \in \mathcal{O}_r$ such that $A \in \mathcal{C}_a(R)$. □

G.8.2 Proof of Theorem 12

First, in order to simplify the computations, we introduce the change of variables $M = SA$ and the function :

$$f_m : M \in \mathbb{R}^{\ell,r} \mapsto \frac{1}{2} \|M\|_F^2 - \|SM\|_*.$$

Note that for any $M \in \mathbb{R}^{\ell,r}$, we have $f_m(M) = f_a(S^{-1}M)$ and $\min_M f_m(M)$ is the form taken by (RRR) if X is the identity and $Y = S$ is a diagonal matrix.

As in section G.7.3, we consider the set $\mathcal{P}$ of permutations $\pi : [\![1; \ell]\!] \to [\![1; \ell]\!]$ such that $\pi(1) < \ldots < \pi(r)$ and simultaneously $\pi(r+1) < \ldots < \pi(\ell)$. For any $\pi \in \mathcal{P}$, we denote :

$$\Pi_\pi := (1_{i=\pi(j)})_{1 \le i \le \ell,\, 1 \le j \le r} \in \mathbb{R}^{\ell,r}. \tag{G.59}$$

With the proposed change of variables, the differentiable critical points of f_m are simply obtained from the critical points of f_a given in Lemma 58.

Lemma 60. *If the values of the diagonal matrix S are strictly decreasing, then the set Ω_m^s of differentiable critical points of f_m is the image by linear transformations from $\mathbb{R}^{r,r}$ to $\mathbb{R}^{\ell,r}$ of $\mathcal{O}_r$:*

$$\Omega_m^s = \{ S\Pi_\pi R |\ \pi \in \mathcal{P}, R \in \mathcal{O}_r \}.$$

The following result describes the eigenvectors of the Hessian of f_m at a critical point $S\Pi_\pi R$. It is proved in Appendix G.11.6. We write $S^2 = \operatorname{diag}(\sigma_1 > \ldots > \sigma_\ell)$ with $\sigma_\ell > 0$. For $1 \le i_0 \le \ell$, $1 \le j_0 \le r$, we denote $E_{i_0,j_0} = e_{i_0} e_{j_0}^T \in \mathbb{R}^{\ell,r}$.

Theorem 61. *Let $S\Pi_\pi R$ be a differentiable critical point of f_m, with $\pi \in \mathcal{P}$ and $R \in \mathcal{O}_r$. Then f_m is twice differentiable at $S\Pi_\pi R$, let $\mathcal{H}_m$ denote its Hessian at $S\Pi_\pi R$.*

- *For $1 \le i < j \le r$, $S^{-1}(E_{\pi(i),j} + E_{\pi(j),i})R$ is an eigenvector of $\mathcal{H}_m$ associated to the eigenvalue 1.*
- *For $1 \le i \le r$, $S^{-1}E_{\pi(i),i}R$ is an eigenvector of $\mathcal{H}_m$ associated to the eigenvalue 1.*

- *For $1 \leq i < j \leq r$, $S(E_{\pi(i),j} - E_{\pi(j),i})R$ is an eigenvector of $\mathcal{H}_m$ associated to the eigenvalue 0.*
- *For $r+1 \leq k \leq \ell$, $1 \leq j \leq r$, $E_{\pi(k),j}R$ is an eigenvector of $\mathcal{H}_m$ associated to the eigenvalue $1 - \frac{\sigma_{\pi(k)}}{\sigma_{\pi(j)}}$.*

Remark 62. *At an optimum $S\tilde{I}R$ of f_m with $R \in \mathcal{O}_r$, the largest eigenvalue of the Hessian is 1 and the smallest positive eigenvalue is $1 - \frac{\sigma_{\pi(r+1)}}{\sigma_{\pi(r)}}$.*

Since we used the change of variables $M = SA$, Theorem 61 can be adapted to the function f_a.

Theorem 63. *Let $\Pi_\pi R$ be a differentiable critical point of f_a, with $\pi \in \mathcal{P}$ and $R \in \mathcal{O}_r$. Then f_a is twice differentiable at $\Pi_\pi R$, let $\mathcal{H}_a$ denote its Hessian at $\Pi_\pi R$.*

- *For $1 \leq i < j \leq r$, $(E_{\pi(i),j} + E_{\pi(j),i})R$ is an eigenvector of $\mathcal{H}_a$ associated to the eigenvalue $(\sigma^{-1}_{\pi(i)} + \sigma^{1}_{\pi(j)})^{-1}$.*
- *For $1 \leq i \leq r$, $E_{\pi(i),i}R$ is an eigenvector of $\mathcal{H}_a$ associated to the eigenvalue $\sigma_{\pi(i)}$.*
- *For $1 \leq i < j \leq r$, $(E_{\pi(i),j} - E_{\pi(j),i})R$ is an eigenvector of $\mathcal{H}_a$ associated to the eigenvalue 0.*
- *For $r+1 \leq k \leq \ell$, $1 \leq j \leq r$, $E_{\pi(k),j}R$ is an eigenvector of $\mathcal{H}_a$ associated to the eigenvalue $\sigma_{\pi(k)}\left(1 - \frac{\sigma_{\pi(k)}}{\sigma_{\pi(j)}}\right)$.*

Proof. Let $\pi \in \mathcal{P}$, $R \in \mathcal{O}_r$ and $\Delta \in \mathbb{R}^{\ell,r}$. Using the change of variables $M = SA$ and denoting $\mathcal{H}_a$ and $\mathcal{H}_m$ the Hessian of respectively f_a at $\Pi_\pi R$ and f_m at $S\Pi_\pi R$, we have the equality :

$$\mathcal{H}_a[\Delta,\, \Delta] = \mathcal{H}_m[S\Delta,\, S\Delta].$$

After normalizing the eigenvectors of $\mathcal{H}_m$ given in Theorem 61, we obtain :

$$\begin{aligned}
\mathcal{H}_a[\Delta R,\, \Delta R] &= \mathcal{H}_m[S\Delta R,\, S\Delta R] \\
&= \sum_{1\leq i<j\leq r} \left\langle (\sigma^{-1}_{\pi(i)} + \sigma^{1}_{\pi(j)})^{-\frac{1}{2}} S^{-1}(E_{\pi(i),j} + E_{\pi(j),i}),\, S\Delta \right\rangle^2 \\
&\quad + \sum_{1\leq i\leq r} \left\langle \sigma^{\frac{1}{2}}_{\pi(i)} S^{-1} E_{\pi(i),i},\, S\Delta \right\rangle^2 \\
&\quad + \sum_{r+1\leq k\leq \ell,\, 1\leq j\leq r} \left(1 - \frac{\sigma_{\pi(k)}}{\sigma_{\pi(j)}}\right) \left\langle E_{\pi(k),j},\, S\Delta \right\rangle^2 \\
&= \sum_{1\leq i<j\leq r} (\sigma^{-1}_{\pi(i)} + \sigma^{1}_{\pi(j)})^{-1} \left\langle E_{\pi(i),j} + E_{\pi(j),i},\, \Delta \right\rangle^2 \\
&\quad + \sum_{1\leq i\leq r} \sigma_{\pi(i)} \left\langle E_{\pi(i),i},\, \Delta \right\rangle^2 \\
&\quad + \sum_{r+1\leq k\leq \ell,\, 1\leq j\leq r} \sigma_{\pi(k)} \left(1 - \frac{\sigma_{\pi(k)}}{\sigma_{\pi(j)}}\right) \left\langle E_{\pi(k),j},\, \Delta \right\rangle^2.
\end{aligned}$$

□

As a direct corollary of Theorem 63, we have the following result.

Corollary 64. *With the notations used in Equation (G.59), an optimum $\tilde{I}R$ of f_a corresponds to the identity permutation $\pi = Id$. At an optimum, the largest eigenvalue of the Hessian $\mathcal{H}_a$ is σ_1 and $\sigma_{\pi(\ell)}\left(1 - \frac{\sigma_{\pi(r+1)}}{\sigma_{\pi(r)}}\right) > 0$ is a lower bound of the positive eigenvalues of $\mathcal{H}_a$.*

The following result is also a straightforward corollary of Theorem 63.

Corollary 65. *All full-rank critical points that are not global minima are strict saddle points* i.e. *the Hessian at these points has a negative eigenvalue.*

Proof. Consider $R \in \mathcal{O}_r$ and a permutation $\pi : [\![1; \ell]\!] \to [\![1; \ell]\!]$ such that $\pi(1) < \ldots < \pi(r)$ and simultaneously $\pi(r+1) < \ldots < \pi(\ell)$ while $\pi \neq Id$. Necessarily, $\pi(r+1) < \pi_r$ and $\sigma_{\pi(r+1)}(1 - \frac{\sigma_{\pi(r+1)}}{\sigma_{\pi(r)}}) < 0$ is an eigenvalue of $\mathcal{H}_a$ at $\Pi_\pi R$ by Theorem 63. □

We can now prove Theorem 12.

Proof of Theorem 12. Consider a minimum $\tilde{I}R$ of f_a with $R \in \mathcal{O}_r$. From Lemma 11, we know that :

$$\mathcal{C}_a(R) = \left\{ \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} R \mid A_1 \in \mathcal{S}_r^+,\ A_2 \in \mathbb{R}^{\ell-r,r} \right\}.$$

We denote the subspace spanned by $\mathcal{C}_a(R)$:

$$\mathcal{E}_R^+ := \text{span } [\mathcal{C}_a(R)] = \left\{ \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} R \mid A_1 \in \mathcal{S}_r,\ A_2 \in \mathbb{R}^{\ell-r,r} \right\},$$

where $\mathcal{S}_r$ is the set of symmetric matrices in $\mathbb{R}^{r,r}$. We know from Theorem 63 that $\mathcal{E}_R^+$ is exactly the subspace spanned by the eigenvectors of the Hessian $\mathcal{H}_{\tilde{I}R}$ of f_a at $\tilde{I}R$ associated to positive eigenvalues. Let $\sigma_{\min} := \sigma_\ell(1 - \frac{\sigma_{r+1}}{\sigma_r})$. As pointed out in Corollary 64, $\sigma_{\min}$ is a lower bound for the positive eigenvalue of the Hessian $\mathcal{H}_{\tilde{I}R}$. Thus, for all $M \in \text{span } (\mathcal{C}_a(R))$, we have :

$$\text{Vec}(M)^T \mathcal{H}_{\tilde{I}R} \text{Vec}(M) \geq \sigma_{\min} \|M\|_F^2,$$

where $\text{Vec}(M) \in \mathbb{R}^{\ell,r}$ is the vectorization of $M \in \mathbb{R}^{\ell,r}$. Given the form of the Hessian for the trace norm in Proposition 6 of [Grave et al., 2011] that is recalled in Proposition 74, the existence of continuous bases for the singular subspaces [Stewart, 2012] of $S^2\tilde{I}$ and the converse of Taylor's Theorem in [Oliver, 1954], we obtain that the Hessian of f_a is continuous at $\tilde{I}R$. Therefore, for any $\gamma < 1 < \delta$, there exists $\alpha > 0$ such that for all $M \in \mathcal{E}_R^+$ and $A \in \mathcal{B}(\tilde{I}R, \alpha) \cap \mathcal{E}_R^+$ where $\mathcal{B}(\tilde{I}R, \alpha)$ is the ball with center $\tilde{I}R$ and radius α, we have

$$\delta\sigma_1 \|M\|_F^2 \geq \text{Vec}(M)^T \mathcal{H}_A \text{Vec}(M) \geq \gamma\sigma_{\min} \|M\|_F^2. \tag{G.60}$$

Consider two elements $M, N \in \mathcal{B}(\tilde{I}R, \alpha)$. The Taylor expansions gives :

$$\begin{aligned} f_a(N) &= f_a(M) + \langle \nabla f_a(M), N - M \rangle \\ &\quad + \frac{1}{2} \int_0^1 \text{Vec}(N - M)^T \mathcal{H}_{tN+(1-t)M} \text{Vec}(N - M) dt \\ &\geq f_a(M) + \langle \nabla f_a(M), N - M \rangle + \frac{\gamma\sigma_{\min}}{2} \|N - M\|_F^2. \end{aligned}$$

This inequality implies that f_a is $\gamma\sigma_{\min}$-strongly convex in $\mathcal{B}(\tilde{I}R, \alpha) \cap \mathcal{E}_R^+$. We conclude by defining a sublevel set $\mathcal{V}_a$ inside $\cup_{R\in\mathcal{O}_r}\mathcal{B}(\tilde{I}R, \alpha)$.

Similarly, we could show from Equation (G.60) that for any A, $A' \in \mathcal{V}_a$ such that $[A, A'] \subset \mathcal{V}_a$, the function f_a has $\delta\sigma_1$-Lipschitz gradients on $[A, A']$. Unfortunately, we can not deduce from this observation that f_a has $\delta\sigma_1$-Lipschitz gradients or is $\delta\sigma_1$-smooth in $\mathcal{V}_a$ since the latter might be nonconvex. However, as in Equation G.15 of Lemma 42, s_1^2 being the largest eigenvalue of S^2, we have for any A, $A' \in \mathbb{R}^{\ell,r}$, such that f_a is differentiable at A,

$$f_a(A') \leq f_a(A) + \langle \nabla f_a(A), A' - A\rangle + \frac{s_1^2}{2}\|A' - A\|_F^2 .$$

Therefore, the function f_a is s_1^2-smooth.

□

Remark 66. *Note that the assumption $s_r > s_{r+1}$ is essential here. In order to highlight its importance, we can give an example to demonstrate that Theorem 12 would not be true if this assumption were not satisfied. Consider $Y = X = I_2 \in \mathbb{R}^{2,2}$ and $r = 1$. Here, the assumptions $s_r > s_{r+1}$ is violated since $s_1 = s_2 = 1$. The cones are $\mathbb{R}_+ \times \mathbb{R}$ and $\mathbb{R}_- \times \mathbb{R}$. The matrix $U = \begin{bmatrix}1\\0\end{bmatrix}$ is an optimum of :*

$$\min_{U\in\mathbb{R}^{2,1}} \frac{1}{2}\|XU\|_F^2 - \left\|Y^TXU\right\|_* = \min_{U\in\mathbb{R}^{2,1}} \frac{1}{2}\|U\|_F^2 - \|U\|_* .$$

However, in the direction $\Delta_\alpha := \begin{bmatrix}0\\\alpha\end{bmatrix}$, there is no strong convexity. Indeed we have :

$$\frac{1}{2}\|X(U + \Delta_\alpha)\|_F^2 = \frac{1}{2}\|U + \Delta_\alpha\|_F^2 = \frac{1}{2}(1 + \alpha^2)$$

and :

$$\left\|Y^TX(U + \Delta_\alpha)\right\|_* = \|U + \Delta_\alpha\|_* = \sqrt{1 + \alpha^2} = 1 + \frac{1}{2}\alpha^2 + o(\alpha^2).$$

By taking the difference of these two equations we prove that there is no second order dependance and consequently no strong convexity in the direction $\begin{bmatrix}0\\1\end{bmatrix}$. It could have been seen directly with Theorem 63 : with $r = 1$, $\ell = 2$, $\pi = Id$, then $E_{\pi(2),1} = E_{2,1} = \begin{bmatrix}0\\1\end{bmatrix}$ is an eigenvector associated to the eigenvalue $\sigma_1(1 - \frac{\sigma_2}{\sigma_1}) = 0$ since $\sigma_1 = \sigma_2 = 1$.

G.8.3 Proof of Corollary 13

Here, we do not assume that X^TX is invertible and prove a more general result. We show that for any $R \in \mathcal{O}_r$ and $N \in \mathbb{R}^{p-m,r}$, the function f restricted to the affine cone $\mathcal{C}(R, N) = \tau(\mathcal{C}_a(R), \mathbb{R}^{m-\ell,r}, N)$, where τ is the function defined in Equation (G.42), is strongly convex in a neighborhood of the optimum $\tau(R, 0, N)$ of f. If we assumed that X^TX is invertible, the proof would be very similar since we would have $m = p$ and the value of $f \circ \tau$ does not depend on N.

Given $R \in \mathcal{O}_r$ and $N \in \mathbb{R}^{p-m,r}$, consider U and U' in the same cone $\mathcal{C}(R,N)$ as $\tau(R,0,N)$. Using the linear change of variables τ, we know that there exists $A, A' \in \mathcal{C}_a(R)$ and $C, C' \in \mathbb{R}^{m-\ell,r}$ such that :

$$\begin{aligned} U &= (X^TX)^{\frac{\dagger}{2}}(PSA + P^\perp C) + K^\perp N \\ &= (X^TX)^{\frac{\dagger}{2}}(PM + P^\perp C) + K^\perp N \text{ with } M = SA, \end{aligned}$$

and similarly $U' = (X^TX)^{\frac{\dagger}{2}}(PM' + P^\perp C') + K^\perp N$ with $M' = SA'$.

We know from Equation (G.43) that :

$$f(U) = \frac{1}{2}\|M\|_F^2 - \|SM\|_* + \frac{1}{2}\|C\|_F^2.$$

In Theorem 61, we have computed the eigenvectors and the eigenvalues of $f_m : M'' \mapsto \frac{1}{2}\|M''\|_F^2 - \|SM''\|_*$ at $S\tilde{I}R$ which is a minimum of f_m. We invoke the same arguments as in the proof of Theorem 12 : given the form of the Hessian for the trace norm in Proposition 6 of [Grave et al., 2011] that is recalled in Proposition 74, the existence of continuous bases for the singular subspaces [Stewart, 2012] of $S^2\tilde{I}$ and the converse of Taylor's Theorem in [Oliver, 1954], we obtain that the Hessian of f_m is continuous at $S\tilde{I}R$. Therefore, for any $\gamma < 1 < \delta$, there exists $\alpha > 0$ such that if $S^{-1}M$, $S^{-1}M' \in \mathcal{B}(\tilde{I}R, \alpha) \cap \mathcal{E}_R^+$ where $\mathcal{B}(\tilde{I}R, \alpha)$ is the ball with center $\tilde{I}R$ and radius α, we have :

$$\begin{aligned} \frac{\gamma}{2}(1 - \frac{s_{r+1}^2}{s_r^2})\|M' - M\|_F^2 &\leq f_m(M) - f_m(M') - \langle \nabla f_m(M'),\, M - M' \rangle \\ &\leq \frac{\delta}{2}\|M' - M\|_F^2, \end{aligned} \tag{G.61}$$

since the smallest positive eigenvalue of the Hessian of f_m at $S\tilde{I}R$ is $1 - \frac{s_{r+1}^2}{s_r^2}$ and the largest is 1.

The variables U and U' being obtained from (M, C, N) and (M', C', N) with a linear transformation, we can define a neighborhood $\mathcal{V}(R,N) \subset \mathcal{C}(R,N)$ of $\tau(\tilde{I}R, 0, N)$ such that U, $U' \in \mathcal{V}(R,N)$ if and only if $S^{-1}M$, $S^{-1}M' \in \mathcal{B}(\tilde{I}R, \alpha) \cap \mathcal{E}_R^+$ and then transfer Equation (G.61) to U and U' :

$$\begin{aligned} &\frac{\gamma}{2}(1 - \frac{s_{r+1}^2}{s_r^2})\left[\|C - C'\|_F^2 + \frac{1}{2}\|M' - M\|_F^2\right] \\ &\leq \frac{1}{2}\|C - C'\|_F^2 + \frac{\gamma}{2}(1 - \frac{s_{r+1}^2}{s_r^2})\|M' - M\|_F^2 \\ &\leq f(U) - f(U') - \langle \nabla f(U'),\, U - U' \rangle. \end{aligned}$$

Also, since $U - U' = (X^TX)^{\frac{\dagger}{2}}(P(M - M') + P^\perp(C - C'))$ we have the following inequality :

$$\|U - U'\|_F^2 \leq d_{\max}^2\left[\,\|M - M'\|_F^2 + \|C - C'\|_F^2\,\right],$$

where $d_{\max}$ is the largest eigenvalue of $(X^TX)^{\frac{\dagger}{2}}$. If X^TX is invertible, $\frac{1}{d_{\max}^2}$ is the smallest eigenvalue of X^TX. Eventually, we obtain :

$$\frac{\gamma}{2d_{\max}^2}(1 - \frac{s_{r+1}^2}{s_r^2})\|U - U'\|_F^2 \leq f(U) - f(U') - \langle \nabla f(U'),\, U - U' \rangle.$$

Setting $\mu := \frac{\gamma}{d_{\max}^2}(1 - \frac{s_{r+1}^2}{s_r^2})$, we have proved that the restriction of f to the affine cone $\mathcal{C}(R, N)$ is μ-strongly convex in the neighborhood $\mathcal{V}(R, N)$ of the optimum $\tau(\tilde{I}R, 0, N)$. We conclude by defining a sublevel set $\mathcal{V}^0 \subset \cup_{R\in\mathcal{O}_r,\, N\in\mathbb{R}^{p-m,r}} \mathcal{V}(R, N)$ of the function f.

The L_X-smoothness of the function f is obtained directly from Equation (G.15) in Fact 42.

Remark 67. *Similarly, we can show from Equation (G.61) that there exists $M > L_X$ such that for any U, $U' \in \mathcal{V}^0$ with $[U, U'] \subset \mathcal{V}^0$, the function f has M-Lipschitz gradients on $[U, U']$, since the Hessian is bounded in $\mathcal{V}^0$. Unfortunately, we cannot deduce from this observation that f has M-Lipschitz gradients in $\mathcal{V}^0$ or is M-smooth in $\mathcal{V}^0$ since the latter might be nonconvex.*

G.8.4 Proof of Corollary 14

To extend to (SRRR) the result that we proved for (RRR), we assume that X^TX is invertible.

Proof of Corollary 14. Let $\mu < \nu_X \left(1\frac{s_r^2}{s_{r+1}^2}\right)$ where ν_X is the samellest eigenvalue of X^TX and $\mathcal{V}^0$ be defined as in Corollary 13. As X^TX is invertible, we know from the orthogonal invariance of $f(U)$ and $\lambda \|U\|_{1,2}$ that for any $R \in \mathcal{O}_r$, a minimum of $F^\lambda(U) = f(U) + \lambda \|U\|_{1,2}$ is attained in each cone $\mathcal{C}(R)$. Theorem 6.4 of Bonnans and Shapiro [1998] guarantees, if its conditions are satisfied, the existence of $\check{\lambda}$ such that for any $R \in \mathcal{O}_r$, the minimum in each cone $\mathcal{C}(R)$ depends continuously on $\lambda \in [0, \check{\lambda})$. The assumptions of the Theorem 6.4 are indeed satisfied and we detail those below :

(a) The objective F^λ of (SRRR) is locally strongly convex on the cone $\mathcal{C}(I_r)$ around the minimum : indeed, the restriction to $\mathcal{C}(I_r)$ of $f : U \mapsto \frac{1}{2}\|XU\|_F^2 - \left\|Y^TXU\right\|_*$ is strongly convex according to Corollary 13 and $\lambda \|U\|_{1,2}$ is convex.

(b) For every fixed λ in some interval $[0, \tilde{\lambda})$, f is locally Lipschitz with a constant that does not depend on λ and the group-Lasso norm is Lipschitz.

(c) The difference $F^\lambda - F^0 = \lambda\|\cdot\|_{1,2}$ is locally Lipschitz with a constant $\sqrt{p}\lambda$ which is $O(\lambda)$.

Thus, according to Theorem 6.4 of Bonnans and Shapiro [1998], there exists $0 < \bar{\lambda} < \check{\lambda}$ such that for any $0 \leq \lambda \leq \bar{\lambda}$, the optimum of (SRRR) in each cone remains in the neighborhood $\mathcal{V}^0$ where f is L_X-smooth and F^λ is μ-strongly convex, with the same constants as f for (RRR). To conclude and obtain Corollary 14, there only remains to define a new open sublevel set $\mathcal{V}^\lambda$ of F^λ inside the sublevel set $\mathcal{V}^0$ of f. □

G.9 Proofs for Section 5.5.3

G.9.1 Proof of Theorem 15

The sequence of inequalities to prove Theorem 15 is the same as in Proof B.1 of Csiba and Richtarik [2017] except for the line search condition that plays the role

of their smoothness condition. Indeed, the result remains true if the function is not smooth as long as the condition (LS) is satisfied. Let $F^{\lambda,*}$ denote the minimum of F^λ. We define define the optimality gap function :

$$\xi : x \mapsto F^\lambda(x) - F^{\lambda,*}.$$

Given $t > 0$ and a point $x \in \mathbb{R}^d$, we have also defined :

$$\begin{aligned} \tilde{f}_{t,x}(x') &:= f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2t} \|x' - x\|_F^2 , \\ \tilde{F}^\lambda_{t,x}(x') &:= \tilde{f}_{t,x}(x') + \lambda h(x'), \end{aligned} \tag{G.62}$$

and x^+ is the unique minimum of the strongly convex function $\tilde{F}^\lambda_{t,x}$.

Proof of Theorem 15. Let $x \in \mathbb{R}^d$ and $t > 0$ such that the condition (LS) is satisfied *i.e.* $\tilde{F}^\lambda_{t,x}(x^+) \geq F^\lambda(x^+)$. We have :

$$\begin{aligned} \xi(x^+) &= F^\lambda(x^+) - F^{\lambda,*} && \text{(G.63)} \\ &\leq \tilde{F}^\lambda_{t,x}(x^+) - F^{\lambda,*} && \text{(G.64)} \\ &= f(x) + \lambda h(x) - F^{\lambda,*} + \langle \nabla f(x), x^+ - x \rangle \\ &+ \frac{1}{2t}\|x^+ - x\|^2 + \lambda h(x^+) - \lambda h(x) && \text{(G.65)} \\ &= \xi(x) + \min_{y \in \mathbb{R}^d} \left[\langle \nabla f(x), y - x \rangle + \frac{1}{2t}\|y - x\|^2 + \lambda h(y) - \lambda h(x) \right] && \text{(G.66)} \\ &= \xi(x) - t\gamma_t(x) && \text{(G.67)} \\ &= \xi(x) \left[1 - t\alpha_t(x) \right]. && \text{(G.68)} \end{aligned}$$

Equation (G.63) follows from the definition of ξ. We have Equation (G.64) since the condition (LS) is satisfied. Equation (G.65) comes from Equation (G.62). Equation (G.66) follows from the definition of x^+, Equation (G.67) from the definition of γ_t and Equation (G.68) from the definitions of α_t and ξ.

□

Remark 68. *A similar result would hold if we used stochastic block coordinate descent like in Lemma 13 of Csiba and Richtarik [2017], the proof would again follow Proof B.1 in Csiba and Richtarik [2017], with the same modification about the condition (LS).*

G.10 Proofs for Section 5.5.4

G.10.1 Proof of Corollary 16

Let μ and $\mathcal{V}^0$ be defined as in Corollary 13. Let $R \in \mathcal{O}_r$ and $U \in \mathcal{C}(R) \cap \mathcal{V}^0$. According to Corollary 13, f is μ-strongly convex on $\mathcal{C}(R) \cap \mathcal{V}^0$. Since the minimal value f^* of f is attained on each cone, let $U^* \in \mathcal{C}(R)$ be an optimum of f. As $\mathcal{C}(R) \cap \mathcal{V}^0$ defines a sublevel set of the restriction of f to $\mathcal{C}(R)$ that is a convex function, it is a convex set. Therefore, the segment $[U^*, U]$ is included in $\mathcal{C}(R) \cap \mathcal{V}^0$.

As a μ-strongly convex function, the restriction $f|_{\mathcal{C}(R)\cap\mathcal{V}^0}$ of f to the convex set $\mathcal{C}(R)\cap\mathcal{V}^0$ satisfies :

$$f|_{\mathcal{C}(R)\cap\mathcal{V}^0}(U^*) \geq f|_{\mathcal{C}(R)\cap\mathcal{V}^0}(U) + \langle \nabla f|_{\mathcal{C}(R)\cap\mathcal{V}^0}(U), U^* - U\rangle + \frac{\mu}{2}\|U^* - U\|_F^2 .$$

Since we have :

$$\langle \nabla f|_{[U,U^*]}(U), U' - U\rangle = \lim_{s\to 0^+} \frac{f(U + s(U' - U)) - f(U)}{s} = \langle \nabla f(U), U' - U\rangle,$$

we obtain :

$$\begin{aligned} f(U) - f^* &\leq \langle \nabla f(U), U - U^*\rangle - \frac{\mu}{2}\|U - U^*\|_F^2 \\ &= \frac{\mu}{2}\left(\left\|\frac{1}{\mu}\nabla f(U)\right\|_F^2 - \left\|U - U^* - \frac{1}{\mu}\nabla f(U)\right\|_F^2\right) \\ &\leq \frac{1}{2\mu}\|\nabla f(U)\|_F^2 . \end{aligned}$$

G.10.2 Proof of Corollary 17

First, we need to introduce the following lemma. It is a light modification of Theorem 15 of Csiba and Richtarik [2017]. Apart from the substitution of the Lipschitz constant with $\frac{1}{t}$, the proof follows Proof B.2 of Csiba and Richtarik [2017].

Lemma 69. *Let $\lambda \geq 0$, $\mu \geq 0$, $\mathcal{C} \subset \mathbb{R}^{p,r}$ a convex set, $f : \mathbb{R}^{p,r} \to \mathbb{R}$ be a differentiable function such that its restriction to $\mathcal{C}$ is μ-strongly convex, $h : \mathbb{R}^{p,r} \to \mathbb{R}$ be a convex function and $F^\lambda = f + \lambda h$. We denote $\bar{f}$, $\bar{h}$ and $\bar{F}^\lambda$ the restrictions of f, h and F^λ to $\mathcal{C}$. $F^{\lambda,*}$ denotes the optimal value of $\bar{F}^\lambda$ in $\mathcal{C}$. Given $U, U' \in \mathcal{C}$ and $t > 0$, we denote :*

$$\tilde{\bar{F}}^\lambda(U') := \bar{f}(U) + \langle \nabla \bar{f}(U), U' - U\rangle + \frac{1}{2t}\|U' - U\|_F^2 + \lambda \bar{h}(U') \tag{G.69}$$

$$\bar{\gamma}_t(U) := -\frac{1}{t}\min_{U'\in\mathcal{C}}\left[\tilde{\bar{F}}^\lambda(U') - \bar{F}^\lambda(U)\right]. \tag{G.70}$$

Let $U \in \mathcal{C}$, $t > 0$ and $U_+ = \operatorname{argmin}_{U'\in\mathcal{C}}\left[\tilde{\bar{F}}^\lambda(U') - \bar{F}^\lambda(U)\right]$. We have :

$$\bar{\gamma}_t(U) \geq \min\left(\frac{1}{2t}, \mu\right)\left[\bar{F}^\lambda(U) - \bar{F}^{\lambda,*}\right].$$

Proof. Let $U \in \mathcal{C}$ such that $\bar{F}^\lambda > \bar{F}^{\lambda,*}$, $t > 0$ and :

$$U_+ = \operatorname*{argmin}_{U'\in\mathcal{C}}\left[\tilde{\bar{F}}^\lambda(U') - \bar{F}^\lambda(U)\right].$$

Then, we have :

$$t\bar{\gamma}_t(U) = -\min_{U' \in \mathcal{C}} \left[\langle \nabla \bar{f}(U), U' - U \rangle + \frac{1}{2t} \|U' - U\|_F^2 + \lambda \bar{h}(U') - \bar{h}(U) \right] \tag{G.71}$$

$$= \bar{F}^\lambda(U) - \min_{U' \in \mathcal{C}} \left[\bar{f}(U) + \langle \nabla \bar{f}(U), U' - U \rangle + \frac{1}{2t} \|U' - U\|_F^2 + \lambda \bar{h}(U') \right]$$

$$\geq \bar{F}^\lambda(U) - \min_{U' \in \mathcal{C}} \left[\bar{f}(U') - \frac{\mu}{2} \|U' - U\|_F^2 + \frac{1}{2t} \|U' - U\|_F^2 + \lambda \bar{h}(U') \right] \tag{G.72}$$

$$= \bar{F}^\lambda(U) - \min_{U' \in \mathcal{C}} \left[\bar{F}^\lambda(U') - \frac{1}{2} \left(\mu - \frac{1}{t} \right) \|U' - U\|_F^2 \right]. \tag{G.73}$$

Equation (G.71) follows from Equation (G.69) and Equation (G.70). Equation (G.72) is due to the μ-strong convexity of $\bar{f}$. We denote $U^* \in \mathcal{C}$ the optimum of $\bar{F}^\lambda$ and for all $U' \in \mathcal{C}$, $\xi(U') := \bar{F}^\lambda(U') - \bar{F}^{\lambda,*}$. Let $0 \leq \delta \leq 1$, setting $U' = U + \delta(U^* - U)$ in Equation (G.73), we obtain :

$$t\bar{\gamma}_t(U) \geq \bar{F}^\lambda(U) - \bar{F}^\lambda(\delta U^* + (1-\delta)U) + \frac{1}{2}\delta^2 \left(\mu - \frac{1}{t} \right) \|U^* - U\|_F^2$$

$$\geq \bar{F}^\lambda(U) - \delta \bar{F}^\lambda(U^*) - (1-\delta)\bar{F}^\lambda(U)$$

$$+ \frac{1}{2} \left[\mu\delta(1-\delta) + \delta^2 \left(\mu - \frac{1}{t} \right) \right] \|U^* - U\|_F^2 \tag{G.74}$$

$$= \delta \left(\xi(U) + \frac{\mu}{2} \|U^* - U\|_F^2 \right) - \frac{\delta^2}{2t} \|U^* - U\|_F^2. \tag{G.75}$$

Equation (G.74) comes from the μ-strong convexity of $\bar{F}^\lambda$. We impose :

$$\delta = \min \left(1, \frac{\xi(U) + \frac{\mu}{2} \|U - U^*\|_F^2}{\frac{1}{t} \|U - U^*\|_F^2} \right). \tag{G.76}$$

Consider the two possible values for δ in Equation (G.76). First, if $\frac{1}{t} \|U - U^*\|_F^2 \leq \xi(U) + \frac{\mu}{2} \|U - U^*\|_F^2$ we have $\delta = 1$ and :

$$\left(\mu - \frac{1}{t} \right) \|U - U^*\|_F^2 \geq \left(\frac{\mu}{2} - \frac{1}{t} \right) \|U - U^*\|_F^2 \geq -\xi(U). \tag{G.77}$$

Combining Equation (G.75) with Equation (G.77) in the case $\delta = 1$, we obtain :

$$t\bar{\gamma}_t(U) \geq \xi(U) + \frac{1}{2} \left(\mu - \frac{1}{t} \right) \|U^* - U\|_F^2 \geq \frac{1}{2}\xi(U). \tag{G.78}$$

Secondly, if $\frac{1}{t} \|U - U^*\|_F^2 \geq \xi(U) + \frac{\mu}{2} \|U - U^*\|_F^2$, we obtain with Equation (G.75) :

$$t\bar{\gamma}_t(U) \geq \frac{\left(\xi(U) + \frac{\mu}{2} \|U^* - U\|_F^2 \right)^2}{\frac{2}{t} \|U^* - U\|_F^2}. \tag{G.79}$$

Therefore, with Equation (G.78) and Equation (G.79), we have :

$$\begin{aligned}\bar{\gamma}_t(U) &\geq \min\left(\frac{1}{2t}\xi(U), \frac{\left(\xi(U) + \frac{\mu}{2}\left\|U^* - U\right\|_F^2\right)^2}{2\left\|U^* - U\right\|_F^2}\right) \\ &\geq \min\left(\frac{1}{2t}\xi(U), \frac{2\xi(U)\mu\left\|U^* - U\right\|_F^2}{2\left\|U^* - U\right\|_F^2}\right) \\ &\geq \min\left(\frac{1}{2t}, \mu\right)\xi(U) = \min\left(\frac{1}{2t}, \mu\right)\left[\bar{F}^\lambda(U) - \bar{F}^{\lambda,*}\right].\end{aligned} \tag{G.80}$$

Equation (G.80) comes from the inequality of arithmetic and geometric means. □

We can now prove Corollary 17. Let $\mathcal{V}^\lambda$ be the sublevel set defined in Corollary 14. Let $R \in \mathcal{O}_r$ and $U \in \mathcal{C}(R) \cap \mathcal{V}^\lambda$. According to Corollary 14, F^λ is μ-strongly convex on $\mathcal{C}(R) \cap \mathcal{V}^\lambda$. Since the minimal value $F^{\lambda,*}$ is attained on each cone, let $U^* \in \mathcal{C}(R)$ be an optimum of $F^{\lambda,*}$. As $\mathcal{C}(R) \cap \mathcal{V}^\lambda$ defines a sublevel set of the restriction of F^λ to $\mathcal{C}(R)$ that is a convex function, it is a convex set. Therefore, the segment $[U^*, U]$ is included in $\mathcal{C}(R) \cap \mathcal{V}^\lambda$.

We define for any $U' \in [U,\, U^*]$ the surrogate $(\tilde{F^\lambda}|_{[U,\,U^*]})_{t,x}(U')$ of the restriction of F^λ to $[U, U^*]$ like in section 5.5.3 :

$$(\tilde{F^\lambda}|_{[U,U^*]})_{t,U}(U') = f(U) + \langle \nabla f|_{[U,U^*]}(U), U' - U\rangle + \frac{1}{2t}\left\|U' - U\right\|_F^2 + \lambda\left\|U'\right\|_{1,2}.$$

From Lemma 69, we obtain the following inequality for any $U' \in [U, U^*]$ such that the condition (LS) is satisfied :

$$\begin{aligned}-\frac{1}{t}\min_{U' \in [U,U^*]}&\left[(\tilde{F^\lambda}|_{[U,U^*]})_{t,U}(U') - F^\lambda|_{[U,U^*]}(U)\right] \\ &\geq \min(\frac{1}{2t}, \mu)\left[F^\lambda(U) - F^{\lambda,*}\right].\end{aligned} \tag{G.81}$$

Since

$$\langle \nabla f|_{[U,U^*]}(U), U' - U\rangle = \lim_{s\to 0^+}\frac{f(U + s(U' - U)) - f(U)}{s} = \langle \nabla f(U), U' - U\rangle,$$

Inequality (G.81) becomes :

$$-\frac{1}{t}\min_{U' \in [U,\,U^*]}\left\{\tilde{F}^\lambda_{t,U}(U') - F^\lambda(U)\right\} \geq \min(\frac{1}{2t}, \mu)\left[F^\lambda(U) - F^{\lambda,*}\right].$$

The minimum over the segment being lower bounded by the minimum over the whole space, we deduce that :

$$\gamma_t(U) \geq \min(\frac{1}{2t}, \mu)\left[F^\lambda(U) - F^{\lambda,*}\right].$$

G.10.3 Proof of Corollary 18

Let $\lambda \geq 0$ and $\mathcal{V}^\lambda$ be a non-empty sublevel set of F^λ such that for all $U \in \mathcal{V}^\lambda$, F^λ satisfies the t-strong *proximal*-PL inequality, as in Corollary 17. Let $k \geq 0$, $t_{k-1} > \frac{\beta}{L_X}$ and $U^k \in \mathcal{V}^\lambda$. If U^{k+1} and t_k are generated as in Algorithm 1 from $U^k \in \mathcal{V}^\lambda$ and t_{k-1} such that the (LS) condition $F^\lambda(U_{k+1}) \leq \tilde{F}^\lambda_{t_k,U_k}(U_{k+1})$ is satisfied, then we know from Fact 43 that the inequality $t_k > \frac{\beta}{L_X}$ is satisfied.

Besides, since we have :

$$F^\lambda(U^{k+1}) \leq \tilde{F}^\lambda_{t_k,U_k}(U^{k+1}) = \min_{U' \in \mathbb{R}^{p,r}} \tilde{F}^\lambda_{t_k,U_k}(U') \leq \tilde{F}^\lambda_{t_k,U_k}(U^k) = F^\lambda(U_k)$$

and $\mathcal{V}^\lambda$ is a sublevel set, it is clear that $U^{k+1} \in \mathcal{V}^\lambda$.

To obtain Equation (18), we can apply Theorem 15 since F^λ satisfies the t_k-strong *proximal*-PL inequality by Corollary 17 with $\alpha(t_k) := \min(\frac{1}{2t_k}, \mu)$:

$$\begin{aligned} F^\lambda(U^{k+1}) - F^{\lambda,*} &\leq [1 - t_k\alpha(t_k)] \left[F^\lambda(U^k) - F^{\lambda,*}\right] \\ &\leq \left[1 - \min(\frac{1}{2}, \mu t_k)\right] \left[F^\lambda(U^k) - F^{\lambda,*}\right] \\ &\leq [1 - \rho] \left[F^\lambda(U^k) - F^{\lambda,*}\right], \end{aligned}$$

where $\rho = \min(\frac{1}{2}, \beta\frac{\mu}{L_X}) \leq \min(\frac{1}{2}, \mu t_k)$.

G.11 Supplementary Results and Proofs

G.11.1 Proof of Lemma 52

First, we prove the following fact.

Fact 70. *If U is a local minimizer of F^λ, then denoting :*

$$V_U \in \operatorname*{argmax}_{V \in \mathbb{R}^{k,r} :\, V^TV = I_r} \langle V, Y^TXU \rangle, \tag{G.82}$$

the matrix $W := UV_U^T \in \mathbb{R}^{p,k}$ has to be a local minimizer of $F_w : W \mapsto \frac{1}{2}\|XW\|_F^2 - \langle Y, XW\rangle + \lambda\|W\|_{1,2}$ among matrices of $\mathbb{R}^{p,k}$ whose rank is smaller than r.

Proof. We prove Fact 70 by contradiction, supposing that $W := UV_U^T$ is not a local minimizer. Without loss of generality, we can assume since F^λ is invariant when its argument is multiplied on the right by an orthogonal matrix that the columns of U are orthogonal. Indeed, if the SVD of U is $R_1\Sigma R_2^T$, we can consider instead $U' = R_1\Sigma$ and modify V_U accordingly. With this assumption, the right singular vectors of $W := UV_U^T$ with V_U defined by Equation (G.82) are exactly the columns of V_U. Since we supposed that W is not a local minimizer, there exists a sequence of matrices $(W_k)_{k\geq 0}$ with rank smaller than r and with limit W such that for each $k \geq 0$, $F_w(W_k) < F_w(W)$. For each $k \geq 0$, let V_k be a matrix with r columns containing at least the right singular vectors of W_k such that $V_k^TV_k = I_r$. In particular, using the continuity of the singular spaces [Stewart, 2012, Theorem V.2.7], we can impose

that the sequence $(V_k)_{k\geq 0}$ has limit V_U. The sequence $(U_k)_{k\geq 0}$ defined for each $k \geq 0$ by $U_k = W_k V_k$ has limit U. For $k \geq 0$, this would mean $W_k = U_k V_k^T$ and :

$$f(U_k) + \lambda \|U_k\|_{1,2} = F^\lambda(U_k) \leq F_w(U_k V_k^T) < F_w(U V_U^T) = f(U) + \lambda \|U\|_{1,2}\,.$$

This would contradict the fact that U is a local minimizer. Therefore $W = UV_U^T$ must be a local minimizer of F_w. □

Proof of Lemma 52. We assume that for any $S \subset \{1, \dots, p\}$ of cardinality at least r, the matrix $X_S^T Y$ is full-rank, where X_S is the matrix formed by keeping the columns of X indexed by S. We prove Lemma 52 by contradiction, assuming that U is a local minimum which has at least r non-zero rows and a rank strictly smaller than r. Again, we denote $V_U \in \operatorname{argmax}_{V^T V = I_r} \langle V, Y^T X U \rangle$ and consider $W := UV_U^T$. First, we write without loss of generality :

$$W = \begin{bmatrix} W_S \\ 0 \end{bmatrix} \text{ , with } |S| \geq r \text{ and } W_S \in \mathbb{R}^{|S|,k} \text{ only has non-zero rows.}$$

Secondly, $\operatorname{rank}(W_S) < r$ since W_S is extracted from W whose rank is smaller than r. According to Fact 70, W is a local minimizer of F_w among matrices with rank smaller than r so for any vectors $u \in \mathbb{R}^p$, $v \in \mathbb{R}^k$, the function $t \mapsto \frac{1}{2}\left\|Y - X(W + tuv^T)\right\|_F^2 + \lambda \left\|W + tuv^T\right\|_{1,2}$ has a minimum at zero. The first-order condition is :

$$u^T X^T (Y - XW) v + \lambda \sum_i u_i z_i^T v = 0,$$

where $u_i \in \mathbb{R}$ and denoting $W_{i,:}$ the i-th row of W, $z_i^T = \frac{W_{i,:}}{\|W_{i,:}\|_2}$ if $W_{i,:}$ is different from zero and z_i has a norm smaller than 1 otherwise. If we impose $v \in \operatorname{Ker} W_S$, we get $Wv = 0$ and $z_i^T v = 0$ for $i \in S$. Therefore we have :

$$u^T X^T Y v + \lambda \sum_{i \notin S} u_i z_i^T v = 0. \tag{G.83}$$

Since Equation (G.83) holds in particular for any $u \in \mathbb{R}^p$ such that $u_i = 0$ when $i \notin S$, we necessarily have for any $v \in \operatorname{Ker} W_S$:

$$X_S^T Y v = 0.$$

In other words, we have $\operatorname{Ker} W_S \subset \operatorname{Ker} X_S^T Y$. This implies that $\dim(\operatorname{Ker} X_S^T Y) \geq \dim(\operatorname{Ker} W_S) > k - r$ since W_S has rank strictly smaller than r. Therefore $X_S^T Y \in \mathbb{R}^{|S|,k}$ has rank strictly smaller than r. This is in contradiction with the assumption in Lemma 52. □

G.11.2 Proof of Lemma 53

Let U^* be a full-rank local minimum of $F^\lambda : U \mapsto \frac{1}{2}\|XU\|_F^2 - \left\|Y^T X U\right\|_* + \lambda \|U\|_{1,2}$. Without loss of generality, we denote S the support of the rows of U^* and we write :

$$U^* = \begin{bmatrix} U_S \\ 0_{p-m,r} \end{bmatrix}$$

where m is the number of non-zero rows of U and $U_S \in \mathbb{R}^{m,r}$. We also denote :

$$X = \begin{bmatrix} X_S & X_{S^c} \end{bmatrix},$$

with $X_S \in \mathbb{R}^{n,m}$ and $X_{S^c} \in \mathbb{R}^{n,p-m}$. Let $V \in \operatorname{argmin}_{V \in \mathbb{R}^{k,r} : V^T V = I_r} \langle Y^T X U^*, V \rangle$ and $G^\lambda : U \mapsto \frac{1}{2}\|XU\|_F^2 - \langle Y^T X U, V \rangle + \lambda \|U\|_{1,2}$. By Fact 33, we have on the one hand $G^\lambda \geq F^\lambda$ and on the other hand $G^\lambda(U^*) = F^\lambda(U^*)$ so U^* is a local minimum of G^λ. The first order conditions restricted to the rows in the set S are :

$$X_S^T X_S U_S - X_S^T Y V + \lambda Z_S = 0, \tag{G.84}$$

where $Z_S := D U_S \in \mathbb{R}^{|S|,r}$ with $D := \operatorname{diag}(\frac{1}{\|U_1^*\|_2}, \ldots, \frac{1}{\|U_m^*\|_2}) \in \mathbb{R}^{|S|,|S|}$ and the norms of the rows of U_S are denoted $\|U_1^*\|_2, \ldots, \|U_m^*\|_2$. In particular, Equation (G.84) implies that :

$$U_S^T \left[X_S^T X_S + D \right] U_S = U_S^T X_S^T Y V.$$

Since we assumed that $|S| \geq r$, the matrix $U_S^T \left[X_S^T X_S + D \right] U_S$ has rank r. Necessarily, $U_S^T X_S^T Y = U^{*T} X^T Y$ also has rank r.

G.11.3 Proof of Lemma 54

Lemma 52 and Lemma 53 combined with Assumption $\mathcal{H}2$ ensure that for any limit point $U \in \bar{\mathcal{U}}$, the matrix $Y^T X U$ is full-rank. Since the set of limit points $\bar{\mathcal{U}}$ is closed and bounded, there exist $\zeta > 0$ and $\delta > 0$ such that for all $U \in \mathbb{R}^{p,r}$, $\operatorname{dist}(U, \bar{\mathcal{U}}) \leq \delta$ implies that the eigenvalues of $Y^T X U$ are lower bounded by ζ, where $\operatorname{dist}(U, \bar{\mathcal{U}})$ is the Euclidean distance between U and the compact set $\bar{\mathcal{U}}$. We denote $\mathcal{K}^\delta := \left\{ U \in \mathbb{R}^{p,r} | \operatorname{dist}(U, \bar{\mathcal{U}}) \leq \delta \right\}$ and $\mathcal{K}^{\frac{\delta}{2}} := \left\{ U \in \mathbb{R}^{p,r} | \operatorname{dist}(U, \bar{\mathcal{U}}) \leq \frac{\delta}{2} \right\}$.

Proposition 6 of [Grave et al., 2011] that is recalled in Proposition 74, describes the Hessian of the trace-norm at full-rank matrices : since for any $U \in \mathcal{K}^\delta$, the eigenvalues of $Y^T X U$ are lower bounded by ζ, there exists $M > 0$ such that the Hessian of f is bounded on $\mathcal{K}^\delta$ by M. Therefore, for any $U, U' \in \mathcal{K}^\delta$ such that $[U, U'] \subset \mathcal{K}^\delta$, we have :

$$\|\nabla f(U) - \nabla f(U')\|_F \leq M \|U - U'\|_F . \tag{G.85}$$

Fact 43 and Lemma 44 ensure that $\lim_{k \to +\infty} \|U_{k+1} - U_k\|_F = 0$ so there exists $k_1 \geq 0$ such that for any $k \geq k_1$, we have $U_k \in \mathcal{K}^{\frac{\delta}{2}}$ and $\|U_{k+1} - U_k\|_F \leq \frac{\delta}{2}$. The triangle inequality implies that $[U_k, U_{k+1}] \subset \mathcal{K}^\delta$. Consequently we have, by Equation (G.85), for all $k \geq k_1$:

$$\|\nabla f(U_k) - \nabla f(U_{k+1})\|_F \leq M \|U_k - U_{k+1}\|_F .$$

G.11.4 Proof of Lemma 57

Let $A \in \mathbb{R}^{\ell,r}$ be a rank deficient matrix and $R_1 D R_2^T$ be a singular value decomposition of the matrix $S^2 A$. Since $S^2 A$ is rank deficient, we can assume that $R_1 \in \mathbb{R}^{\ell,r-1}$, $D \in \mathbb{R}^{r-1,r-1}$ and $R_2 \in \mathbb{R}^{r,r-1}$. Up to a multiplication on the right by an orthogonal matrix, we can assume, using the orthogonal invariance of f_a, that :

$$S^2 A = R_1 D \begin{bmatrix} I_{r-1} & 0_{r-1} \end{bmatrix}, \text{ where } I_{r-1} \in \mathbb{R}^{r-1,r-1}, 0_{r-1} \in \mathbb{R}^{r-1}.$$

Let $e_r \in \mathbb{R}^r$ be the vector whose components are 0 except for the last one that is 1. Let $t \in \mathbb{R}$ and $\tilde{a} \in \mathbb{R}^\ell$ be a unit-norm vector such that $S^2\tilde{a}$ is orthogonal to the columns of R_1 and therefore to the columns of S^2A. We have $\left\|\tilde{a}e_r^T\right\|_F = 1$. On the one hand, we can separate the Frobenius norm of $S(A + t\tilde{a}e_r^T)$ as follows :

$$\frac{1}{2}\|S(A+t\tilde{a}e_r^T)\|_F^2 = \frac{1}{2}\|SA\|_F^2 + \frac{1}{2}t^2\|S\tilde{a}e_r^T\|_F^2 = \frac{1}{2}\|SA\|_F^2 + \frac{1}{2}t^2\|S\tilde{a}\|_F^2 = \frac{1}{2}\|SA\|_F^2 + o(t).$$

On the other hand, for any $t \neq 0$, a singular value decomposition of $S^2(A + t\tilde{a}e_r^T)$ is :

$$S^2(A + t\tilde{a}e_r^T) = \begin{bmatrix} R_1 & \frac{S^2\tilde{a}}{\|S^2\tilde{a}\|_F} \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & |t|\|S^2\tilde{a}\|_F \end{bmatrix} \begin{bmatrix} I_{r-1} & 0_{r-1} \\ 0_{r-1}^T & \frac{t}{|t|} \end{bmatrix}.$$

We can therefore easily compute the trace norm of $S^2(A + t\tilde{a}e_r^T)$:

$$\|S^2(A + t\tilde{a}e_r^T)\|_* = \|S^2A\|_* + |t|\|S^2\tilde{a}\|_F \geq \|S^2A\|_* + |t|s_\ell^2,$$

where s_ℓ is the smallest eigenvalue of S. So finally, we obtain :

$$f_a(A + t\tilde{a}e_r^T) \leq f_a(A) - s_\ell^2|t| + o(t).$$

G.11.5 Proof of Lemma 58

Proof. Let A be a critical point of $f_a : A \mapsto \frac{1}{2}\|SA\|_F^2 - \|S^2A\|_*$ and denote $V_A \in \operatorname{argmax}_{V \in \mathbb{R}^{\ell,r};\, V^TV = I_r} \langle S, SAV^T \rangle$. We know from Lemma 57 that A is full-rank and applying Danskin's Theorem [Danskin, 1967], we have :

$$\nabla f_a(A) = S^2(A - V_A) = 0. \tag{G.86}$$

Besides, writing $\Pi\Sigma R$ the singular value decomposition of S^2A with $\Pi \in \mathbb{R}^{\ell,r}$ a matrix whose columns are orthogonal, $\Sigma \in \mathbb{R}^{r,r}$ a diagonal matrix whose entries are denoted $\sigma_1, \ldots, \sigma_\ell$ and $R \in \mathbb{R}^{r,r}$ an orthogonal matrix, we know that $A = V_A$ from Equation (G.86) and that $V_A = \Pi R$ by Fact 35. Therefore, we have :

$$\begin{aligned} & S^2A = S^2\Pi R \\ \Rightarrow \quad & \Pi\Sigma R = S^2\Pi R \quad \text{since } \Pi\Sigma R \text{ is the SVD of } S^2A, \\ \Rightarrow \quad & \Pi\Sigma = S^2\Pi \quad \text{since } RR^T = I_r. \end{aligned} \tag{G.87}$$

Let $i \in [\![1, r]\!]$, $w := (w_1, \ldots, w_\ell)^T$ be the i-th column of Π. Equation (G.87) implies that :

$$\begin{aligned} \sigma_i w &= \begin{bmatrix} s_1^2 w_1 \\ \vdots \\ s_\ell^2 w_\ell \end{bmatrix}, \\ \Rightarrow \quad & \forall j \in [\![1, r]\!],\ (\sigma_i - s_j^2) w_j = 0. \end{aligned}$$

Since we assumed that $s_1, \ldots, s_\ell$ are all different, only one w_j can be different from zero and must be 1 since w has norm 1. Given that the columns of the matrix Π are orthogonal and contain only one nonzero coefficient, up to a permutation of its columns, the matrix Π has the form given in Lemma 58.

With Lemma 55, we know that A is not a local maximum of f_a. If $A = \tilde{I}R$, we have proved in Lemma 9 that A is a global minimum. Now assume that $A = \Pi_\pi R$, with π and Π_π as in Equation (G.55) and that there exists $i \in \{1, \dots, r\}$ such that $\pi(i) > i$ and for all $i' < i$, $\pi(i') = i'$. We have :

$$\frac{1}{2}\left\|S(A + te_ie_i^TR)\right\|_F^2 = \frac{1}{2}\|SA\|_F^2 + \frac{t^2}{2}s_i^2. \tag{G.88}$$

Since the Frobenius norm of the i-th column of $S^2(A + te_ie_i^TR)$ is $\sqrt{s_{\pi(i)}^4 + t^2 s_i^4}$ and the columns of $S^2(A + te_ie_i^TR)$ have disjoint supports, we also have :

$$\begin{aligned}\left\|S^2(A + te_ie_i^TR)\right\|_* &= \left\|S^2A\right\|_* - s_{\pi(i)}^2 + \sqrt{s_{\pi(i)}^4 + t^2s_i^4} \\ &= \left\|S^2A\right\|_* - s_{\pi(i)}^2 + s_{\pi(i)}^2\left(1 + \frac{t^2}{2}\frac{s_i^4}{s_{\pi(i)}^4}\right) + O(t^4). \end{aligned}\tag{G.89}$$

Combining Equation (G.88) with Equation (G.89), we obtain :

$$f_a(A + te_ie_i^TR) - f_a(A) = \frac{t^2s_i^2}{2}\left(1 - \frac{s_i^2}{s_{\pi(i)}^2}\right) + O(t^4).$$

Since we have assumed that $\pi(i) > i$ and the eigenvalues of the matrix S are strictly decreasing, we have $\left(1 - \frac{s_i^2}{s_{\pi(i)}^2}\right) < 0$ and A is not a local minimum. □

G.11.6 Proof of Theorem 61

As in section G.7.3, we consider a permutation $\pi : [\![1; \ell]\!] \to [\![1; \ell]\!]$ such that simultaneously $\pi(1) < \dots < \pi(r)$ and $\pi(r+1) < \dots < \pi(\ell)$. We denote :

$$\Pi_\pi := (1_{i=\pi(j)})_{1\le i\le \ell,\, 1\le j\le r} \in \mathbb{R}^{\ell,r},$$

and define for $i_0 \in [\![1, \ell]\!]$ and $j_0 \in [\![1, r]\!]$:

$$E_{i_0,j_0} = (1_{i=i_0,\, j=j_0})_{1\le i\le \ell,\, 1\le j\le r} \in \mathbb{R}^{\ell,r}.$$

We want to compute the Hessian $\mathcal{H}_m$ of $f_m : M \mapsto \frac{1}{2}\|M\|_F^2 - \|SM\|_*$ at the matrix $M = S\Pi_\pi R$. It is well defined according to Proposition 6 in [Grave et al., 2011] since $SM = S^2\Pi_\pi R$ is full-rank. We recall this result below in Proposition 74. In order to introduce the different eigenvectors of the Hessian of f_m, we need the singular value decomposition and the polar decomposition of SM. Since $M = S\Pi_\pi R$ and $S^2\Pi_\pi = \Pi_\pi \mathrm{diag}(s_{\pi(1)}^2, \dots, s_{\pi(r)}^2)$, a singular value decomposition of SM is given by :

$$SM = \Pi_\pi \mathrm{diag}(s_{\pi(1)}^2, \dots, s_{\pi(r)}^2)R, \quad \Pi_\pi^T\Pi_\pi = I_r \quad \text{and} \quad R^TR = I_r.$$

We have $s_{\pi(1)}^2 > \dots > s_{\pi(r)}^2$ because we assumed $s_1 > \dots > s_\ell > 0$ and $\pi(1) < \dots < \pi(r)$. Defining $V = \Pi_\pi R \in \mathbb{R}^{\ell,r}$ and $K = R^T\mathrm{diag}(s_{\pi(1)}^2, \dots, s_{\pi(r)}^2)R \in \mathbb{R}^{r,r}$, we obtain the polar decomposition of SM :

$$SM = VK, \quad V^TV = I_r \quad \text{and} \quad K \in \mathcal{S}_r^{++}, \tag{G.90}$$

with $\mathcal{S}_r^{++}$ the set of positive-definite matrices in $\mathbb{R}^{r,r}$. We also denote $\mathcal{S}_r = \{H \in \mathbb{R}^{r,r} \mid H^T = H\}$ the set of symmetric matrices in $\mathbb{R}^{r,r}$.

First we focus on a set of directions where the restriction of f_m is exactly a quadratic strongly convex function.

Fact 71. *The restriction of $M' \mapsto \|SM'\|_*$ to the affine subspace $\{M + S^{-2}MH \mid H \in \mathcal{S}_r\}$ is linear in a neighborhood of M, its Hessian at M is zero. Consequently, the Hessian of $f_m : M \mapsto \frac{1}{2}\|M\|_F^2 - \|SM\|_*$ restricted to the subspace $T_{\mathcal{K}} := \{S^{-2}MH \mid H \in \mathcal{S}_r\}$ is exactly the identity. A basis for $T_{\mathcal{K}}$ is the concatenation of $(S^{-1}(E_{\pi(i),j} + E_{\pi(j),i})R)_{1\leq i<j\leq r}$ with $(S^{-1}E_{\pi(i),i}R)_{1\leq i\leq r}$.*

Proof. For any matrix $\tilde{M}$ such that the polar decomposition of $S\tilde{M}$ has the form VB with $B \in \mathcal{S}_r^+$, we have $\left\|S\tilde{M}\right\|_* = \langle S\tilde{M}, V\rangle$. Indeed, if QDQ^T is a singular value decomposition of B with $Q \in \mathbb{R}^{r,r}$, $Q^TQ = I_r$ and $D \in \mathbb{R}^{r,r}$ a diagonal matrix, then $(VQ)DQ^T$ is a singular value decomposition of VB. Using Fact 33 and Fact 35, we have :

$$\left\|S\tilde{M}\right\|_* = \langle S\tilde{M}, (VQ)Q^T\rangle = \langle S\tilde{M}, V\rangle.$$

Consequently, we have :

$$f_m(\tilde{M}) = \frac{1}{2}\left\|\tilde{M}\right\|_F^2 - \langle S\tilde{M}, V\rangle.$$

In particular, for any $\Delta = S^{-1}VH$ with $H \in \mathcal{S}_r$ such that $K + H \in \mathcal{S}_r^+$, we have $M + \Delta = S^{-1}V(K + H)$ since $M = S^{-1}VK$ according to Equation (G.90) and :

$$f_m(M + \Delta) = \frac{1}{2}\|M + \Delta\|_F^2 - \langle S(M + \Delta), V\rangle.$$

Therefore, the Hessian of $\Delta \mapsto f_m(M + \Delta)$ restricted to the subspace $T_{\mathcal{K}} := \{S^{-1}VH,\, H \in \mathcal{S}_r\}$ is locally the identity. Note that $S^{-1}V = S^{-2}S\Pi_\pi R = S^{-2}M$ since $V = \Pi_\pi R$ and $M = S\Pi_\pi R$ so :

$$T_{\mathcal{K}} = \{S^{-2}MH,\, H \in \mathcal{S}_r\}.$$

We can also use $M = S\Pi_\pi R$ to write :

$$\begin{aligned} T_{\mathcal{K}} &= \{S^{-1}\Pi_\pi RH,\, H \in \mathcal{S}_r\} \\ &= \{S^{-1}\Pi_\pi HR,\, H \in \mathcal{S}_r\}. \end{aligned} \tag{G.91}$$

For Equation (G.91), we have used the fact that for any orthogonal matrix $R \in \mathbb{R}^{r,r}$, the application $H \mapsto R^THR$ is an automorphism of $\mathcal{S}_r$. We then obtain a basis for $T_{\mathcal{K}}$ using the fact that the concatenation of $(E_{i,j} + E_{j,i})_{1\leq i<j\leq r}$ with $(E_{i,i})_{1\leq i\leq r}$ is a basis of $\mathcal{S}_r$ and for any $1 \leq i,\, j \leq r$, $\Pi_\pi E_{i,j} = E_{\pi(i),j}$. □

Secondly, the invariance of f_m when its argument is multiplied on the right by an orthogonal matrix gives a set of directions included in the kernel of the Hessian.

Fact 72. *The subspace $T_{\mathcal{R}} := \{MT \mid T^T = -T,\, T \in \mathbb{R}^{r,r}\}$ is included in the Kernel of the Hessian of f_m at $M = S\Pi_\pi R$. Additionally, $T_{\mathcal{K}} \oplus^\perp T_{\mathcal{R}} = \{MF \mid F \in \mathbb{R}^{r,r}\}$ and a basis for $T_{\mathcal{R}}$ is $(S(E_{\pi(i),j} - E_{\pi(j),i})R)_{1\leq i<j\leq r}$.*

Proof. Since M is a critical point of f_m which is invariant when its argument is multiplied on the right by an orthogonal matrix, then by [Li et al., 2016, Theorem 2], the subspace that is tangent to the manifold $\{MR' \mid R' \in \mathcal{O}_r\}$ is included in the Kernel of the Hessian of f_m at M. In Example 4, Li et al. [2016] show that this subspace is exactly $T_{\mathcal{R}} := \{MT \mid T \in \mathbb{R}^{r,r}, T^T = -T\}$. Since $M = S\Pi_\pi R$ and the set of antisymmetric matrices of $\mathbb{R}^{r,r}$ can be written $\{R^T T R \mid T \in \mathbb{R}^{r,r}, T^T = -T\}$, a basis for $T_{\mathcal{R}}$ is $(S(E_{\pi(i),j} - E_{\pi(j),i})R)_{1 \leq i < j \leq r}$.

To show that $\{MF \mid F \in \mathbb{R}^{r,r}\}$ can be decomposed with the given orthogonal sum, it is first important to notice that :

$$\begin{aligned} T_{\mathcal{K}} &= \{S^{-1}VH \mid H \in \mathcal{S}_r\} \\ &= \{MK^{-1}H \mid H \in \mathcal{S}_r\}. \end{aligned} \tag{G.92}$$

We have used Equation (G.90) to obtain Equation (G.92). It is then sufficient to notice that both $T_{\mathcal{K}} = \{MK^{-1}H \mid H \in \mathcal{S}_r\}$ and $T_{\mathcal{R}} = \{MT \mid T^T = -T, T \in \mathbb{R}^{r,r}\}$ are included in $\{MF, F \in \mathbb{R}^{r,r}\}$, they are also orthogonal given the bases that we have introduced and finally, their dimensions are respectively $\frac{r(r+1)}{2}$ and $\frac{r(r-1)}{2}$ since M is full-rank so their sum must be equal to $\{MF \mid F \in \mathbb{R}^{r,r}\}$ which is of dimension r^2. □

What remains to study is the eigenvectors and the corresponding eigenvalues of the Hessian of f_m at M in the subspace that is orthogonal to $\{MF \mid F \in \mathbb{R}^{r,r}\}$.

Fact 73. *For $r+1 \leq k \leq \ell$ and $1 \leq j \leq r$, the matrix $E_{\pi(k),j}R$ is an eigenvector of the Hessian of f_m restricted to the subspace $T_{V^\perp} := \{C \in \mathbb{R}^{\ell,r} \mid M^T C = 0\}$ and the corresponding eigenvalue is $1 - \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}}$.*

To prove Fact 73, we use the following result.

Proposition 74. *[Grave et al., 2011, Proposition 6] Let $\ell \geq r$, $N \in \mathbb{R}^{\ell,r}$ be a full-rank matrix and $W\Sigma Z^T \in \mathbb{R}^{\ell,r}$ be its singular value decomposition, with $W \in \mathbb{R}^{\ell,r}$, $W^T W = I_r$, $\Sigma = diag(\sigma_1 \geq \ldots \geq \sigma_r) \in \mathbb{R}^{r,r}$ with $\sigma_r > 0$, $Z \in \mathbb{R}^{r,r}$ and $Z^T Z = I_r$. Let $W_0 \in \mathbb{R}^{\ell,\ell-r}$ such that $W_0^T W_0 = I_{\ell-r}$ and $W^T W_0 = 0$. We denote $(w_i)_{1 \leq i \leq r}$ the columns of W, $(z_j)_{1 \leq j \leq r}$ the columns of Z and $(w_k)_{r+1 \leq k \leq \ell}$ the columns of W_0. For any $\Delta \in \mathbb{R}^{\ell,r}$, we have :*

$$\begin{aligned} \|N + \Delta\|_* = \|N\|_* &+ \langle WZ^T, \Delta \rangle \\ &+ \frac{1}{2} \sum_{1 \leq i \leq r,\, 1 \leq j \leq r} \frac{(w_i^T \Delta z_j - w_j^T \Delta z_i)^2}{2(\sigma_i + \sigma_j)} \\ &+ \frac{1}{2} \sum_{r+1 \leq k \leq \ell,\, 1 \leq j \leq r} \frac{(w_k^T \Delta z_j)^2}{\sigma_j} + o(\|\Delta\|_F^2). \end{aligned} \tag{G.93}$$

Proof of Fact 73. Given a perturbation ΔR of the matrix M, we have

$$\|SM + S\Delta R\|_* = \left\|S^2 \Pi_\pi R + S\Delta R\right\|_*, \tag{G.94}$$

$$= \left\|S^2 \Pi_\pi + S\Delta\right\|_* \tag{G.95}$$

Equation (G.94) comes from $M = S\Pi_\pi R$ and we have Equation (G.95) since the trace-norm is orthogonal invariant. Thus, we apply Proposition 74 for a perturbation

$S\Delta$ of the matrix $S^2\Pi_\pi$ whose singular value decomposition is $\Pi_\pi \mathrm{diag}(s^2_{\pi(1)} > \ldots > s^2_{\pi(r)})$. With the notations of Proposition 74, this corresponds to $W = \Pi_\pi$, $\Sigma = \mathrm{diag}(s^2_{\pi(1)} > \ldots > s^2_{\pi(r)})$ and $Z = I_r$. Let $W_0 \in \mathbb{R}^{\ell,\ell-r}$ be the matrix whose columns w_k are the $e_{\pi(k)}$ for $r+1 \leq k \leq \ell$, then $W_0^T W_0 = I_{\ell-r}$ and $W^T W_0 = 0$. We have :

$$\begin{aligned}
\left\|S^2\Pi_\pi + S\Delta\right\|_* &= \left\|S^2\Pi_\pi\right\|_* + \langle WZ^T, S\Delta\rangle \\
&\quad + \frac{1}{2}\sum_{1\leq i\leq r,\, 1\leq j\leq r} \frac{(w_i^T S\Delta z_j - w_j^T S\Delta z_i)^2}{2(s^2_{\pi(i)} + s^2_{\pi(j)})} \\
&\quad + \frac{1}{2}\sum_{r+1\leq k\leq \ell,\, 1\leq j\leq r} \frac{(w_k^T S\Delta z_j)^2}{s^2_{\pi(j)}} + o(\|\Delta\|_F^2) \\
&= \left\|S^2\Pi_\pi\right\|_* + \langle \Pi_\pi, S\Delta\rangle \\
&\quad + \frac{1}{2}\sum_{1\leq i\leq r,\, 1\leq j\leq r} \frac{(s_{\pi(i)} e^T_{\pi(i)}\Delta e_j - s_{\pi(j)} e^T_{\pi(j)}\Delta e_i)^2}{2(s^2_{\pi(i)} + s^2_{\pi(j)})} \\
&\quad + \frac{1}{2}\sum_{r+1\leq k\leq \ell,\, 1\leq j\leq r} \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}}(e^T_{\pi(k)}\Delta z_j)^2 + o(\|\Delta\|_F^2).
\end{aligned}$$

Note that in the first sum, Δ only intervenes through a product with the transpose of an element $e_{\pi(i)}$ that belongs to Im M. Since we already studied the effect of the Hessian on the subspace $\{MF \mid F \in \mathbb{R}^{r,r}\}$ in Fact 71 and Fact 72, we focus on the effect of the Hessian in the orthogonal subspace that is described in the second sum. Given a perturbation Δ of the form $W_0 F Z^T$ with $F \in \mathbb{R}^{\ell-r,r}$, we have on the one hand :

$$\begin{aligned}
\left\|S^2\Pi_\pi + S\Delta\right\|_* &= \left\|S^2\Pi_\pi\right\|_* + \langle \Pi_\pi, S\Delta\rangle \\
&\quad + \frac{1}{2}\sum_{r+1\leq k\leq \ell,\, 1\leq j\leq r} \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}}(e^T_{\pi(k)} W_0 F Z^T z_j)^2 + o(\|\Delta\|_F^2) \\
&= \left\|S^2\Pi_\pi\right\|_* + \langle \Pi_\pi, S\Delta\rangle \\
&\quad + \frac{1}{2}\sum_{r+1\leq k\leq \ell,\, 1\leq j\leq r} \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}} F^2_{k-r,j} + o(\|\Delta\|_F^2). \qquad \text{(G.96)}
\end{aligned}$$

Equation (G.96) comes from $e^T_{\pi(k)} W_0 = (1_{i=k-r})^T_{1\leq i\leq \ell-r}$ and $Z^T z_j = (1_{i=j})_{1\leq i\leq r}$.

On the other hand, we have :

$$\begin{aligned}
\frac{1}{2}\|M + \Delta R\|_F^2 &= \frac{1}{2}\|M\|_F^2 + \frac{1}{2}\|\Delta R\|_F^2 + \langle M,\, \Delta R\rangle \\
&= \frac{1}{2}\|M\|_F^2 + \frac{1}{2}\|\Delta\|_F^2 + \langle S\Pi_\pi R,\, \Delta R\rangle \qquad \text{(G.97)} \\
&= \frac{1}{2}\|M\|_F^2 + \frac{1}{2}\left\|W_0 F Z^T\right\|_F^2 + \langle S\Pi_\pi,\, \Delta\rangle \qquad \text{(G.98)} \\
&= \frac{1}{2}\|M\|_F^2 + \frac{1}{2}\sum_{1\leq i\leq \ell-r,\, 1\leq j\leq r} F^2_{i,j} + \langle S\Pi_\pi,\, \Delta\rangle. \qquad \text{(G.99)}
\end{aligned}$$

Equation (G.97) follows from $M = S\Pi_\pi R$, Equation (G.98) from $\Delta = W_0 F Z^T R$ and Equation (G.99) from $W_0^T W_0 = I_{\ell-r}$ and $Z = I_r$. Combining Equation (G.96) with Equation (G.99), we obtain for $\Delta = W_0 F Z^T$:

$$f_m(M + \Delta R) = f_m(M) + \frac{1}{2} \sum_{r+1 \leq k \leq \ell,\, 1 \leq j \leq r} \left(1 - \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}}\right) F^2_{k-r,j} + o(\|\Delta\|_F^2).$$

Since $W_0 \in \mathbb{R}^{\ell,\ell-r}$ is the matrix whose columns are the $e_{\pi(k)}$ for $r+1 \leq k \leq \ell$ and $Z = I_r$, we obtain the last eigenvectors of the Hessian of f_m : for $r+1 \leq k \leq \ell$ and $1 \leq j \leq r$, the matrix $E_{\pi(k),j} R$ is an eigenvector associated to the eigenvalue $1 - \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}}$. □

Remark 75. *Note that we could have directly used Equation (G.93) to prove simultaneously Fact 71, Fact 72 and Fact 73 but we believe that the proposed analysis helps understanding the structure of the eigenspaces.*

Eventually, we have proved that the Hessian of f_m at M is block diagonal on the three orthogonal subspaces :

- $T_{\mathcal{K}} := \{S^{-2} M H \mid H \in \mathcal{S}_r\}$ where the eigenvalues are all equal to 1.
- $T_{\mathcal{R}} := \{MT \mid T^T = -T\}$ where the eigenvalues are all 0.
- $T_{V^\perp} := \{W_0 C \mid C \in \mathbb{R}^{\ell-r,r}\}$ where the eigenvalues are the $1 - \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}}$ for $r+1 \leq k \leq \ell$, $1 \leq j \leq r$.

We summarize the eigenvectors of the Hessian of $f_m : M' \mapsto \frac{1}{2}\|M'\|_F^2 - \|SM'\|_*$ at $M = S\Pi_\pi R$ in the table below.

Eigenvectors and Eigenvalues of the Hessian of $f_m : M' \mapsto \frac{1}{2}\|M'\|_F^2 - \|SM'\|_*$ at $M = \Pi_\pi R$			
Indices	Number of elements	Eigenvectors	Eigenvalues
$1 \leq i \leq r$	r	$S^{-1} E_{\pi(i),i} R$	1
$1 \leq i < j \leq r$	$\frac{r(r-1)}{2}$	$S^{-1}(E_{\pi(i),j} + E_{\pi(j),i}) R$	1
$1 \leq i < j \leq r$	$\frac{r(r-1)}{2}$	$S(E_{\pi(i),j} - E_{\pi(j),i}) R$	0
$r+1 \leq k \leq \ell$, $1 \leq j \leq r$	$r(\ell - r)$	$E_{\pi(k),j} R$	$1 - \frac{s^2_{\pi(k)}}{s^2_{\pi(j)}}$

G.12 KŁ with exponent $\frac{1}{2}$

As announced at the end of Section 5.5.4, we show in Section G.12.1 that the geometric structure leveraged in Corollary 17 can be used to prove that F^λ has the KŁ property with exponent $\frac{1}{2}$ near the set of optima. While in the core of the article, we proposed a direct proof of Corollary 18 based on Corollary 17 and Theorem 15, we present in Section G.12.2 an application of the framework developed by Csiba and Richtarik [2017] to show that the KŁ property with exponent $\frac{1}{2}$ (instead of the PŁ inequality) also leads to linear convergence. The proofs appear simpler than the ones encountered in [Attouch and Bolte, 2009; Attouch et al., 2013; Chouzenoux et al., 2014; Frankel et al., 2015] as the algorithms considered in these papers are more general while we restrain our study to the proximal gradient algorithm with line search.

G.12.1 KŁ-$\frac{1}{2}$ on cones for (RRR / SRRR)

We assume that X^TX is invertible. Let span $\mathcal{C}(I_r)$ be the subspace spanned by $\mathcal{C}(I_r) = \tau(\mathcal{C}_a(I_r), \mathbb{R}^{p-\ell,r})$ with τ defined in Equation (5.6) and $\mathcal{C}_a(I_r)$ defined in Equation (5.8). Let $F^\lambda_{I_r}$ be the restriction of F^λ to $\mathcal{C}(I_r)$: it is defined for any $U \in \operatorname{span} \mathcal{C}(I_r)$ as $F^\lambda_{I_r}(U) = F^\lambda(U)$ if $U \in \mathcal{C}(I_r)$ and $F^\lambda_{I_r}(U) = +\infty$ otherwise. From the structure described in Corollary 14, and since τ is a linear invertible change of variables, we know that $F^\lambda_{I_r} \circ \tau$ is strongly convex in a neighborhood of $(\tilde{I}, 0_{p-\ell,r}) = (\begin{bmatrix} I_r \\ 0_{\ell-r,r} \end{bmatrix}, 0_{p-\ell,r})$ included in $\mathcal{C}_a(I_r)$.

Fact 76. *Let F be a proper lower semi-continuous function. If F is μ-strongly convex in a set $\mathcal{V} \subset \mathbb{R}^d$ then given $x^* \in \mathcal{V}$, F has the Kurdyka-Łojasiewicz property at $x^* \in \operatorname{dom} \partial F$ with exponent $1/2$: there exist $\eta > 0$ and a neighborhood $\mathcal{U}$ of x^* such that for all $x \in \mathcal{U} \cap \{y \mid F(x^*) < F(y) < F(x^*) + \eta\}$, we have :*

$$\frac{c}{\sqrt{F(x) - F(x^*)}} \operatorname{dist}(0, \partial F(x)) \geq 1. \tag{G.100}$$

Proof. Let $x^* \in \mathcal{V}$. First, if $0 \notin \partial F(x^*)$, then by Lemma 2 of Attouch et al. [2010], there is $c > 0$ and a neighborhood $\mathcal{U}$ of x^* such that for any $x \in \mathcal{U}$, we have :

$$\operatorname{dist}(0, \partial F(x)) \geq \frac{1}{c} \quad \text{and} \quad F(x) - F(x^*) \leq 1,$$

so Equation (G.100) holds for any $x \in \mathcal{U}$.

Secondly, assume that $0 \in \partial F(x^*)$. Let $x \in \mathcal{V}$ such that $F(x) > F(x^*)$ and

$v \in \partial F(x)$. Since F is μ-strongly convex, we have :

$$\begin{aligned} F(x) - F(x^*) &\leq \langle v, x - x^* \rangle - \frac{\mu}{2} \|x - x^*\|^2 \\ &= \frac{\mu}{2} \left[\frac{1}{\mu^2} \|v\|^2 - \frac{1}{\mu^2} \|v\|^2 + 2\langle \frac{1}{\mu} v, x - x^* \rangle - \|x - x^*\|^2 \right] \\ &= \frac{\mu}{2} \left[\frac{1}{\mu^2} \|v\|^2 - \left\| x - x^* - \frac{1}{\mu} v \right\|^2 \right] \\ &\leq \frac{1}{2\mu} \|v\|^2 . \end{aligned}$$

Therefore, we obtain Equation (G.100) with $c = \frac{1}{\sqrt{2\mu}}$:

$$\frac{1}{\sqrt{2\mu}} \frac{1}{\sqrt{F(x) - F(x^*)}} \operatorname{dist}(0, \partial F(x)) \geq 1.$$

□

Since $F^{\lambda}_{I_r}$ is strongly convex, it is a KŁ-$\frac{1}{2}$ function by Fact 76. This is key to apply the following result.

Theorem 77. *[Theorem 3.2 in Li and Pong, 2017] Consider $a \geq b \geq 1$, $g : \mathbb{R}^b \to \mathbb{R}$ a proper closed function and $h : \mathbb{R}^a \to \mathbb{R}^b$ a continuously differentiable mapping. Suppose in addition that g is a KŁ function with exponent $\alpha \in [0, 1)$ and the Jacobian $Jh(\bar{x}) \in \mathbb{R}^{b,a}$ is a surjective mapping at some $\bar{x} \in \operatorname{dom} g \circ h$. Then $g \circ h$ has the KŁ property at $\bar{x}$ with exponent α.*

Let $I_{p-\ell,r} : \mathbb{R}^{p-\ell,r} \mapsto \mathbb{R}^{p-\ell,r}$ be the identity function and $\bar{\sigma} : \mathbb{R}^{\ell,r} \to \mathcal{C}(I_r)$ be the function defined for full-rank matrices based on the polar decomposition :

$$\bar{\sigma} : \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} R \in \mathbb{R}^{\ell,r} \mapsto \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},$$

where $B_1 \in \mathcal{S}_r^{++}$, $R \in \mathcal{O}_r$. This definition is correct as the polar decomposition of a full-rank matrix B_1 is unique. Given the orthogonal invariance of F^{λ} and $F^{\lambda} \circ \tau$, we have :

$$F^{\lambda} = F^{\lambda}_{I_r} \circ \tau \circ (\bar{\sigma}, I_{p-\ell,r}) \circ \tau^{-1}.$$

Before applying Theorem 77 with $g = F^{\lambda}_{I_r} \circ \tau$ and $h = \bar{\sigma} \circ \tau^{-1}$, we first have to prove that its assumptions are satisfied. Clearly, the Jacobian of τ, τ^{-1} and $I_{p-\ell,r}$ are surjective since these are linear invertible functions.

Proposition 78. *Let $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \in \mathbb{R}^{\ell,r}$ such that $A_1 \in \mathbb{R}^{r,r}$ is a square invertible matrix and $A_2 \in \mathbb{R}^{\ell-r,r}$. The Jacobian $J\bar{\sigma}(A)$ is a surjective mapping.*

Proof. Thanks to the polar decomposition, we know that there exists $B_1 \in \mathcal{S}_r^{++}$, $B_2 \in \mathbb{R}^{\ell-r,r}$ and $R \in \mathcal{O}_r$ such that :

$$A = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} R.$$

Consequently, we have $\bar{\sigma}(A) = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$.

Also, given $\Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} \in \mathbb{R}^{\ell,r}$ such that $\Delta_1 \in \mathbb{R}^{r,r}$, $\Delta_2 \in \mathbb{R}^{\ell-r,r}$ and $A + \Delta \in \mathcal{C}_R$, we can write $\begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} R$ where $M_1 \in \mathbb{R}^{r,r}$ is a symmetric matrix, $M_2 \in \mathbb{R}^{\ell-r,r}$ and :

$$\bar{\sigma}(A+\Delta) = (A+\Delta)R^T = \bar{\sigma}(A) + \Delta R^T = \bar{\sigma}(A) + \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}.$$

Therefore, we can identify the differential of $\bar{\sigma}$ on the set of matrices $\begin{bmatrix} M_1 \\ M_2 \end{bmatrix} R$ with M_1 symmetric with the linear application $M \mapsto MR^T$. The surjectivity of this differential is obvious. □

Corollary 79. *Let $0 \leq \lambda < \bar{\lambda}$ and a sublevel set $\mathcal{V}^\lambda$ be defined as in Corollary 14. The function F^λ has the KŁ property with exponent $1/2$ in the sublevel set $\mathcal{V}^\lambda$.*

Proof. According to Fact 76, $F^\lambda_{I_r} \circ \tau$ is a KŁ-$\frac{1}{2}$ function around its optimum since it is locally strongly convex. Consequently, $F^\lambda = \left[F^\lambda_{I_r} \circ \tau\right] \circ \left[(\bar{\sigma}, I_{p-\ell,r}) \circ \tau^{-1}\right]$ is the composition in the sublevel set $\mathcal{V}^\lambda$ of a KŁ-$\frac{1}{2}$ function with a smooth function that has a surjective Jacobian mapping, according to Proposition 78. We deduce with Theorem 77 that F^λ has the KŁ property with exponent $\frac{1}{2}$ in $\mathcal{V}^\lambda$. □

G.12.2 From KŁ with exponent $\frac{1}{2}$ to (t-strong *proximal* PŁ)

Here, we prove that the KŁ-$\frac{1}{2}$ property in $\mathcal{V}^\lambda$ for the function F^λ of SRRR leads to linear convergence for Algorithm 1. This result differs from Theorem 15 of Csiba and Richtarik [2017] for which they assumed strong-convexity instead of the KŁ property with exponent $\frac{1}{2}$.

As in Theorem 51 we make the assumptions $\mathcal{H}2$ and $\mathcal{H}3$ in this section so that we can use Lemma 54. Indeed, we need these extra assumptions because although the function f we consider for SRRR is L_X-smooth with L_X the largest eigenvalue of X^TX, it may not have Lipschitz gradients in the entire sublevel set defined in Corollary 14, mainly because the latter is not convex.

We denote for any $U \in \mathcal{V}^\lambda$, and $t > 0$:

$$\tilde{F}^\lambda_{t,U}(U') := f(U) + \langle \nabla f(U), U' - U \rangle + \frac{1}{2t}\|U' - U\|^2 + \lambda \|U'\|_{1,2}, \tag{G.101}$$

$$\gamma_t(U) := -\frac{1}{t} \min_{U' \in \mathbb{R}^{p,r}} \left[\tilde{F}^\lambda_{t,U}(U') - F^\lambda(U)\right]. \tag{G.102}$$

Before obtaining in Proposition 81 a result similar to the (t-strong *proximal* PŁ) property, we first need to introduce the following result. It is highly similar to Lemma 44 but is is adapted to the present context.

Lemma 80. *Let $U \in \mathcal{V}^\lambda$ and $U_+ := \operatorname{argmin}_{U' \in \mathbb{R}^{p,r}} \left[\tilde{F}^\lambda_{t,U}(U') - F^\lambda(U)\right]$. There is a subgradient s_{U_+} of $\|\cdot\|_{1,2}$ at U_+ such that :*

$$U_+ - U = -t\left(\nabla f(U) + \lambda s_{U_+}\right), \tag{G.103}$$

$$\gamma_t(U) \geq \frac{1}{2}\|\nabla f(U) + \lambda s_{U_+}\|^2. \tag{G.104}$$

Proof. Equation (G.103) is a direct consequence of the first-order optimal conditions for Problem (G.102). We also have :

$$\begin{aligned}
\tilde{F}^{\lambda}_{t,U}(U_+) - F^{\lambda}(U) &= f(U) + \langle \nabla f(U), U_+ - U\rangle + \frac{1}{2t}\|U_+ - U\|_F^2 \\
&\quad + \lambda \|U_+\|_{1,2} - f(U) - \lambda \|U\|_{1,2} && \text{(G.105)} \\
&= \langle \nabla f(U) + \lambda s_{U_+}, U_+ - U\rangle + \frac{t}{2}\|\nabla f(U) + \lambda s_{U_+}\|_F^2 \\
&\quad + \lambda \left[\|U_+\|_{1,2} + \langle s_{U_+}, U - U_+\rangle - \|U\|_{1,2}\right] && \text{(G.106)} \\
&\leq -\frac{t}{2}\|\nabla f(U) + \lambda s_{U_+}\|_F^2. && \text{(G.107)}
\end{aligned}$$

In Equation (G.105), we simply use Equation (G.101). Equation (G.106) follows from Equation (G.103). Equation (G.107) follows again from Equation (G.103) and from the convexity of $\|\cdot\|_{1,2}$. Therefore, we have :

$$\gamma_t(U) \geq \frac{1}{2}\|\nabla f(U) + \lambda s_{U_+}\|^2.$$

□

Proposition 81. *Let $k_1 \geq 0$ be defined as in Lemma 54, $k \geq k_1$ and assume that $U_k \in \mathcal{V}^{\lambda} \backslash \Omega^*$. Let $U_{k+1} = \operatorname{argmin}_{U' \in \mathbb{R}^{p,r}} \left[\tilde{F}^{\lambda}_{t_k,U_k}(U') - F^{\lambda}(U_k)\right]$. We have :*

$$c^2(1 + (Mt_k)^2)\gamma_{t_k}(U_k) \geq F^{\lambda}(U_{k+1}) - F^{\lambda,*}. \tag{G.108}$$

Proof. We know from Lemma 80 that there exists a subgradient $s_{U_{k+1}}$ of $U' \mapsto \|U'\|_{1,2}$ at U_{k+1} such that :

$$U_{k+1} = U_k - t_k \left[\nabla f(U_k) + \lambda s_{U_{k+1}}\right]. \tag{G.109}$$

We have :

$$\begin{aligned}
\|\nabla f(U_{k+1}) + \lambda s_{U_{k+1}}\|^2 &\leq 2\|\nabla f(U_k) + \lambda s_{U_{k+1}}\|^2 + 2\|\nabla f(U_k) - \nabla f(U_{k+1})\|^2 && \text{(G.110)} \\
&\leq 2\|\nabla f(U_k) + \lambda s_{U_{k+1}}\|^2 + 2M^2\|U_k - U_{k+1}\|^2 && \text{(G.111)} \\
&\leq 2\|\nabla f(U_k) + \lambda s_{U_{k+1}}\|^2 \\
&\quad + 2(Mt_k)^2\|\nabla f(U_k) + \lambda s_{U_{k+1}}\|^2 && \text{(G.112)} \\
&\leq 2(1 + (Mt_k)^2)\|\nabla f(U_k) + \lambda s_{U_{k+1}}\|^2 && \text{(G.113)} \\
&\leq 4(1 + (Mt_k)^2)\gamma_t(U_k) && \text{(G.114)}
\end{aligned}$$

We obtain Equation (G.110) using the triangle inequality and the inequality of arithmetic and geometric means. Equation (G.111) is due to Lemma 54. We have Equation (G.112) thanks to Equation (G.103). Equation (G.113) follows from Equation (G.104) in Lemma 80.

Since $\|\nabla f(U_{k+1}) + \lambda s_{U_{k+1}}\|^2$ is an upper bound of $\operatorname{dist}(0, \partial F^{\lambda}(U_{k+1}))^2$, Equation G.114 implies that :

$$\operatorname{dist}(0, \partial F^{\lambda}(U_{k+1}))^2 \leq 4(1 + (Mt_k)^2)\gamma_{t_k}(U_k).$$

Besides, we know from Corollary 79 that there exists $c > 0$ such that for any $U' \in \mathcal{V}^\lambda$, the function F^λ satisfies the inequality :

$$\text{dist}\left(0, \partial F^\lambda(U')\right) \geq \frac{2}{c}\sqrt{F^\lambda(U') - F^{\lambda,*}}. \tag{G.115}$$

The element U_{k+1} being in the sublevel set $\mathcal{V}^\lambda$ since $F^\lambda(U_{k+1}) \leq F^\lambda(U_k)$ and $U_k \in \mathcal{V}^\lambda$, we finally obtain with Equation (G.115) :

$$(1 + (Mt_k)^2)\gamma_t(U_k) \geq \frac{1}{c^2}(F^\lambda(U_{k+1}) - F^{\lambda,*})$$

□

Remark 82. *Note that Equation* (t-strong *proximal* PŁ) *in Section 5.5.3, that is to say in the PŁ framework, can be written :*

$$\gamma_t(U) \geq c_1[F^\lambda(U) - F^{\lambda,*}] \quad \textit{with } c_1 > 0,$$

while Equation G.108, in the KŁ framework, can be written :

$$\gamma_t(U) \geq c_2[F^\lambda(U_+) - F^{\lambda,*}] \quad \textit{with } c_2 > 0.$$

The right term depends either on U or U_+ and this is the main reason for the differences found in the computations between the two frameworks.

Proposition 81 finally leads to local linear convergence, as encountered in Proposition 5.1 of Li and Pong [2017] for batch proximal gradient descent. As in Proposition 5.1 of Li and Pong [2017], we have to use an upper bound $d > 0$ on the step size t while this was not necessary when we used the Polyak-Łojasiewicz inequality instead of the Kurdyka-Łojasiewciz inequality. We denote :

$$\xi : U' \mapsto F^\lambda(U') - F^{\lambda,*}.$$

Proposition 83. *Let $k_1 \geq 0$ be defined as in Lemma 54. Assume that there is $d > 0$ such that for any $k \geq k_1$, we have $t_k \leq d$. There is $0 < \rho < 1$ such that for any $k \geq k_1$, if $U_k \in \mathcal{V}^\lambda \backslash \Omega^*$, then we have :*

$$\xi(U_{k+1}) \leq (1 - \rho)\xi(U_k).$$

Therefore, the convergence of Algorithm 1 is locally linear.

Proof. Let $k \geq k_1$, $U_k \in \mathcal{V}^\lambda$ and :

$$U_{k+1} = \underset{U' \in \mathbb{R}^{p,r}}{\text{argmin}} \left[\tilde{F}^\lambda_{t_k, U_k}(U') - F^\lambda(U_k)\right].$$

First, from Equation (G.108) in Proposition 81, we have :

$$\frac{\gamma_{t_k}(U_k)}{\xi(U_{k+1})} \geq \frac{1}{c^2(1 + (Mt_k)^2)}. \tag{G.116}$$

Secondly, we have :

$$\xi(U_{k+1}) \leq \xi(U_k) - t_k \gamma_{t_k}(U_k) \tag{G.117}$$
$$\leq \xi(U_k) - t_k \frac{\gamma_t(U_k)}{\xi(U_{k+1})})\xi(U_{k+1})$$
$$\leq \xi(U_k) - t_k \frac{1}{c^2(1+(Mt_k)^2)}\xi(U_{k+1}) \tag{G.118}$$
$$\leq \xi(U_k) - t_k \frac{1}{c^2(1+(Md)^2)}\xi(U_{k+1}). \tag{G.119}$$

Equation (G.117) comes from Fact 41. Equation (G.118) follows from Equation G.116 and Equation (G.119) follows from the assumption $t_k \leq d$. Consequently, we have :

$$\begin{aligned}\xi(U_{k+1}) &\leq \frac{1}{1+\frac{t_k}{c^2(1+(Md)^2)}}\xi(U_k) \\ &\leq \frac{1}{1+\frac{\beta}{c^2 L_X(1+(Md)^2)}}\xi(U_k) \qquad (G.120) \\ &\leq (1-\rho)\xi(U_k) \quad \text{with } \rho := 1 - \frac{1}{1+\frac{\beta}{c^2 L_X(1+(Md)^2)}}.\end{aligned}$$

We have Inequality (G.120) since $t_k > \frac{\beta}{L_X}$ for k sufficiently large by Fact 43. □

Proposition 83 finally leads to local linear convergence for the proximal-gradient algorithm applied to (SRRR). The proof in this section is different from the core of the article since we used KŁ inequalities instead of PŁ inequalities.

G.13 Additional details and results on the experiments

G.13.1 Algorithm of Park et al. [2016]

To evaluate the performance of Algorithm 1 for RRR, we compare it with the algorithm proposed in Park et al. [2016], which minimizes the biconvex formulation of Problem (5.1) *i.e.* with the form of Equation (5.4) . To avoid the scaling issue due to the invariance of the objective by any transformation $(U, V) \mapsto (UC, VC^{-T})$ where C is a square invertible matrix, the formulation that they propose has an additional regularizer $(U, V) \mapsto \frac{1}{4}\|U^T U - V^T V\|_F^2$ which does not change the optimal value of the function. With this differentiable function, simultaneous gradient descent in U and V is feasible. However, this regularization scheme is not applicable if a group-Lasso penalty is added, because the latter is not compatible with imposing the constraint $U^T U = V^T V$ at the optima.

G.13.2 Different values of the correlation coefficient ρ

Given that the choice of the correlation coefficient ρ has a strong impact on the running time, we report in Figure G.2 and Figure G.3 additional results for different

values of the parameter ρ. Apart from this modification, we test the algorithms with the same setting as in Section 5.6. This change corresponds to modifying the correlation between the columns of the design matrix X. Although the speed of the algorithms decreases when ρ increases, the relative order of the methods remains the same.

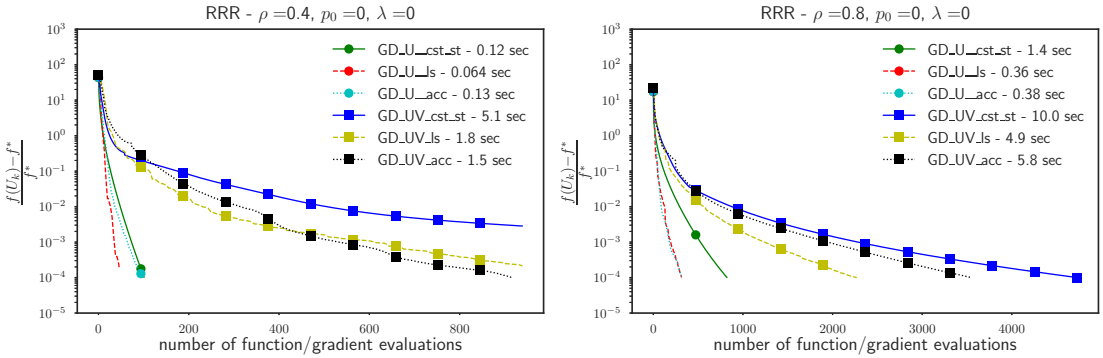


FIGURE G.2: *(Left)* RRR : $\rho = 0.4$. *(Right)* RRR : $\rho = 0.8$. Times reported are times to reach a gap of 10^{-4}

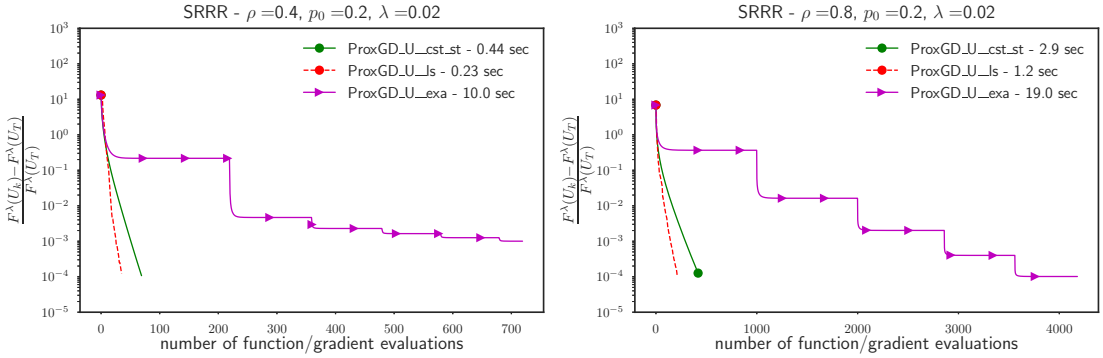


FIGURE G.3: *(Left)* SRRR : $\rho = 0.4$. *(Right)* SRRR : $\rho = 0.8$. Times reported are times to reach a gap of 10^{-4}.

G.13.3 Different sparsity scenarios

To assess the quality of the algorithm when the proportion of zero rows in W_0 varies, we present Figure G.4 where the proportion p_0 of zero rows is respectively 0.5 and 0.8, that is W_0 has 50% and 80% of zero rows.

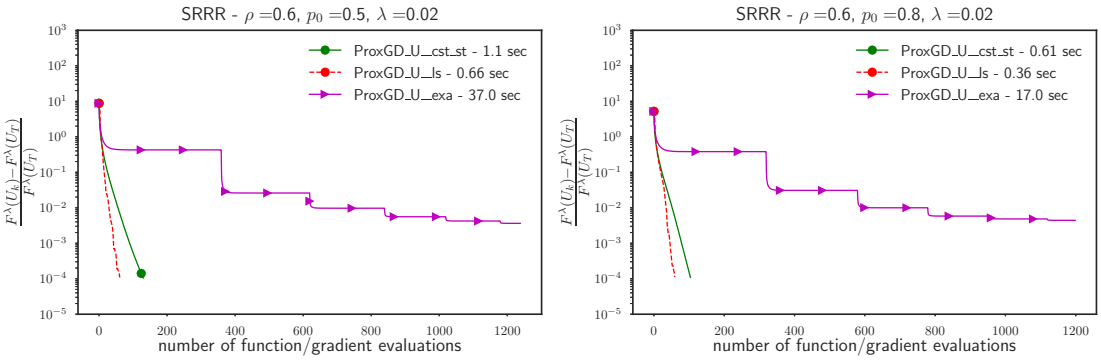


FIGURE G.4: *(Left)* SRRR : $\rho = 0.6$, $p_0 = 0.5$ and $\lambda = 0.02$. *(Right)* SRRR : $\rho = 0.6$, $p_0 = 0.8$ and $\lambda = 0.02$. Times reported are times to reach a gap of 10^{-4}.

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331):1248–1255.

Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853.

Ando, R. K. and Zhang, T. (2007). Two-view feature generation model for semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 25–32. ACM.

Antoniadis, A., Brosat, X., Cugliari, J., and Poggi, J.-M. (2012). Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. application à la consommation d'électricité.

Antoniadis, A., Brosat, X., Cugliari, J., and Poggi, J.-M. (2014). Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité. *Journal de la Société Française de Statistique*, 155(2):202–219.

Antoniadis, A., Paparoditis, E., and Sapatinas, T. (2006). A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):837–857.

Arnold, V. I. (1957). On functions of three variables. In *Doklady Akademii Nauk*, volume 114, pages 679–681. Russian Academy of Sciences.

Attouch, H. and Bolte, J. (2009). On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457.

Attouch, H., Bolte, J., and Svaiter, B. F. (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1-2):91–129.

Auder, B., Cugliari, J., Goude, Y., and Poggi, J.-M. (2018). Scalable clustering of individual electrical curves for profiling and bottom-up forecasting. *Energies*, 11(7):1893.

Avellaneda, M. and Boyer-Olson, D. (2002). Reconstruction of volatility: pricing index options by the steepest descent approximation. *Courant Institute-NYU Working Paper*.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106.

Bakin, S. et al. (1999). Adaptive regression and model selection in data mining problems.

Bakker, B. and Heskes, T. (2003). Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research*, 4(May):83–99.

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis : Learning from examples without local minima. *Neural networks*, 2(1):53–58.

Barbier, T. (2017). *Modélisation de la consommation électrique à partir de grandes masses de données pour la simulation des alternatives énergétiques du futur*. PhD thesis, MINES ParisTech.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

Ben-David, S. and Schuller, R. (2003). Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016). Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881.

Binev, P., Cohen, A., Dahmen, W., and DeVore, R. (2007). Universal algorithms for learning theory. part ii: Piecewise polynomial functions. *Constructive approximation*, 26(2):127–152.

Bolte, J., Daniilidis, A., and Lewis, A. (2007). The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223.

Bolte, J., Sabach, S., and Teboulle, M. (2014). Proximal alternating linearized minimization or nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494.

Bonnans, J. F. and Shapiro, A. (1998). Optimization problems with perturbations : A guided tour. *SIAM review*, 40(2):228–264.

Boumal, N., Voroninski, V., and Bandeira, A. (2016). The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Bruhns, A., Deurveilher, G., and Roy, J.-S. (2005). A non linear regression model for mid-term load forecasting and improvements in seasonality. In *Proceedings of the 15th Power Systems Computation Conference*, pages 22–26. Citeseer.

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, pages 1282–1309.

Bunea, F., She, Y., Wegkamp, M. H., et al. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5):2359–2388.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chen, B.-J., Chang, M.-W., et al. (2004). Load forecasting using support vector machines: A study on eunite competition 2001. *IEEE transactions on power systems*, 19(4):1821–1830.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4.

Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2013). Modeling and forecasting daily electricity load curves: a hybrid approach. *Journal of the American Statistical Association*, 108(501):7–21.

Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2015). Modelling and forecasting daily electricity load via curve linear regression. In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, pages 35–54. Springer.

Chouzenoux, E., Pesquet, J.-C., and Repetti, A. (2014). Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132.

Cont, R. and Deguest, R. (2013). Equity correlations implied by index options : estimation and model uncertainty analysis. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 23(3):496–530.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403.

CRE (2019). L'électricité, comment ça marche ? http://modules-pedagogiques.cre.fr/m1/index.html. Last accessed on August 05, 2019.

Cros, S. and Pinson, P. (2018). Prévision météorologique pour les énergies renouvelables. *La Météorologie*, 2018(100 Spécial Anniversaire 25 ans).

Csiba, D. and Richtarik, P. (2017). Global convergence of arbitrary-block gradient methods for generalized Polyak-Łojasiewicz functions. *arXiv preprint arXiv:1709.03014*.

Cugliari, J., Goude, Y., and Poggi, J.-M. (2016). Disaggregated electricity forecasting using wavelet-based clustering of individual consumers. In *2016 IEEE International Energy Conference (ENERGYCON)*, pages 1–6. IEEE.

Danskin, J. M. (1967). *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media.

Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

Devaine, M., Gaillard, P., Goude, Y., and Stoltz, G. (2013). Forecasting electricity consumption by aggregating specialized experts. *Machine Learning*, 90(2):231–260.

Dordonnat, V., Koopman, S. J., Ooms, M., Dessertaine, A., and Collet, J. (2008). An hourly periodic state space model for modelling french national electricity load. *International Journal of Forecasting*, 24(4):566–587.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. (2017). Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077.

Dubois, B., Delmas, J.-F., and Obozinski, G. (2019). Fast algorithms for sparse reduced-rank regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2415–2424.

Duchemin, Q. (2018). Modèles de clustering pour la prévision de la consommation électrique. Report, LIGM, UMR 8049, École des Ponts, UPEM, ESIEE Paris, CNRS, UPE, Champs-sur-Marne, France.

Dudek, G. (2015). Short-term load forecasting using random forests. In *Intelligent Systems' 2014*, pages 821–828. Springer.

Dumont, M., Marée, R., Wehenkel, L., and Geurts, P. (2009). Fast multi-class image annotation with random subwindows and multiple output randomized trees. In *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 196–203.

Durrleman, V. and El Karoui, N. (2008). Coupling smiles. *Quantitative Finance*, 8(6):573–590.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, pages 89–102.

Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637.

Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, S. and Hyndman, R. J. (2011). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1):134–141.

Fisher, R. A. (1919). Xv.-the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

Forster, B. (2011). Splines and multiresolution analysis. In *Handbook of Mathematical Methods in Imaging*, pages 1231–1270. Springer.

Frankel, P., Garrigos, G., and Peypouquet, J. (2015). Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Friedman, J. H. et al. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823.

Friedman, J. H. and Stuetzle, W. (1982). Smoothing of scatterplots. Technical report, Stanford University CA Project Orion.

Gaillard, P. and Goude, Y. (2015). Forecasting electricity consumption by aggregating experts; how to design a good set of experts. In *Modeling and stochastic learning for forecasting in high dimensions*, pages 95–115. Springer.

Gaillard, P., Goude, Y., and Nedellec, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, 32(3):1038–1050.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44.

Ge, R., Jin, C., and Zheng, Y. (2017). No spurious local minima in nonconvex low rank problems: a unified geometric analysis. *arXiv preprint arXiv:1704.00708*.

Ge, R., Lee, J. D., and Ma, T. (2016). Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981.

Gelfand, A. E. and Dalal, S. R. (1990). A note on overdispersed exponential families. *Biometrika*, 77(1):55–64.

Gelman, A. et al. (2005). Analysis of variance-why it is more important than ever. *The annals of statistics*, 33(1):1–53.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. *Trabajos de estadística y de investigación operativa*, 31(1):489.

Goude, Y., Nedellec, R., and Kong, N. (2013). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE transactions on smart grid*, 5(1):440–446.

Grave, E., Obozinski, G. R., and Bach, F. R. (2011). Trace lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems*, pages 2187–2195.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):149–170.

Gross, C. W. and Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, 9(3):233–254.

Hahn, H., Meyer-Nieberg, S., and Pickl, S. (2009). Electric load forecasting methods: Tools for decision making. *European journal of operational research*, 199(3):902–907.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Heskes, T. (2000). Empirical Bayes for learning to learn.

Hippert, H. S., Pedreira, C. E., and Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1):44–55.

Hofmann, T. and Puzicha, J. (1999). Latent class models for collaborative filtering. In *IJCAI*, volume 99.

Hong, T. and Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938.

Hong, T., Pinson, P., and Fan, S. (2014). Global energy forecasting competition 2012.

Huang, S.-J. and Shih, K.-R. (2003). Short-term load forecasting via ARMA model identification including non-Gaussian process considerations. *IEEE Transactions on power systems*, 18(2):673–679.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589.

IEA (2016). Energy policies of iea countries - France. Report, International Energy Agency.

IEA (2019). France energy balance Sankey diagram. https://www.iea.org/sankey/#?c=France&s=Balance - International Energy Agency. Last accessed on August 25, 2019.

Intrator, N. and Edelman, S. (1996). Making a low-dimensional representation suitable for diverse tasks. In *Learning to learn*, pages 135–157. Springer.

Jacob, L., Vert, J.-P., and Bach, F. R. (2009). Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems*, pages 745–752.

Jain, P., Jin, C., Kakade, S., and Netrapalli, P. (2017). Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488.

Jian, L., Tao, H., and Meng, Y. (2018). Real-time anomaly detection for very short-term load forecasting. *Journal of Modern Power Systems and Clean Energy*, 6(2):235–243.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017). How to escape saddle points efficiently. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). Scipy: Open source scientific tools for Python.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145.

Jourdain, B. and Sbai, M. (2012). Coupling index and stocks. *Quantitative Finance*, 12(5):805–818.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726.

Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594.

Khamaru, K. and Wainwright, M. (2018). Convergence guarantees for a class of non-convex and non-smooth optimization problems. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2601–2610.

Khotanzad, A., Afkhami-Rohani, R., Lu, T.-L., Abaye, A., Davis, M., and Maratukulam, D. J. (1997). ANNSTLF-a neural-network-based electric load forecasting system. *IEEE Transactions on Neural networks*, 8(4):835–846.

Kiartzis, S., Bakirtzis, A., and Petridis, V. (1995). Short-term load forecasting using neural networks. *Electric Power Systems Research*, 33(1):1–6.

Kim, S.-J. and Giannakis, G. B. (2013). Load forecasting via low rank plus sparse matrix factorization. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1682–1686. IEEE.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*, volume 114, pages 953–956. Russian Academy of Sciences.

Kolter, J. Z. and Ferreira, J. (2011). A large-scale study on predicting and contextualizing building energy usage. In *Twenty-fifth AAAI conference on artificial intelligence*.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409.

Kumar, A. and Daume III, H. (2012). Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*.

Kyriakides, E. and Polycarpou, M. (2007). Short term electric load forecasting: A tutorial. In *Trends in Neural Computation*, pages 391–418. Springer.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. (2017). First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*.

Li, G. and Pong, T. K. (2017). Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, pages 1–34.

Li, H. and Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In *Advances in neural information processing systems*, pages 379–387.

Li, Q., Zhu, Z., and Tang, G. (2017). Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*.

Li, Q., Zhu, Z., and Tang, G. (2018). The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*.

Li, X., Wang, Z., Lu, J., Arora, R., Haupt, J., Liu, H., and Zhao, T. (2016). Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology)*, 61(2):381–400.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Ma, Z. and Sun, T. (2014). Adaptive sparse reduced-rank regression. *arXiv*, 1403.

Maclaurin, D., Duvenaud, D., and Adams, R. P. (2015). Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, volume 238.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mammen, E., van de Geer, S., et al. (1997). Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413.

Maurer, A. (2009). Transfer bounds for linear feature learning. *Machine learning*, 75(3):327–350.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*, volume 2. Chapman and Hall, London.

Mei, J., De Castro, Y., Goude, Y., Azaïs, J.-M., and Hébrail, G. (2018). Nonnegative matrix factorization with side information for time series recovery and prediction. *IEEE Transactions on Knowledge and Data Engineering*, 31(3):493–506.

Mei, J., De Castro, Y., Goude, Y., and Hébrail, G. (2017). Nonnegative matrix factorization for time series recovery from a few temporal aggregates. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2382–2390. JMLR. org.

Mei, J., Goude, Y., Hebrail, G., and Kong, N. (2016). Spatial estimation of electricity consumption using socio-demographic information. In *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pages 753–757. IEEE.

Messner, J. W. and Pinson, P. (2018). Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. *International Journal of Forecasting*.

Mougeot, M., Picard, D., Lefieux, V., and Maillard-Teyssier, L. (2015). Forecasting intra day load curves using sparse functional regression. In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, pages 161–181. Springer.

Mukherjee, A., Chen, K., Wang, N., and Zhu, J. (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2):457–477.

Muñoz, A., Sánchez-Úbeda, E. F., Cruz, A., and Marín, J. (2010). Short-term forecasting in power systems: a guided tour. In *Handbook of power systems II*, pages 129–160. Springer.

Nagbe, K., Cugliari, J., and Jacques, J. (2018). Short-term electricity demand forecasting using a functional state space model. *Energies*, 11(5):1120.

Nagbe, K., Cugliari, J., Thebault, A., and Jacques, J. (2017). Prévision de génération d'électricité à partir de sources renouvelables.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nikolova, M. and Tan, P. (2017). Alternating proximal gradient descent for nonconvex regularised problems with multiconvex coupling terms. *HAL-01492846, 2017*.

Nowicka-Zagrajek, J. and Weron, R. (2002). Modeling electricity loads in california: Arma models with hyperbolic noise. *Signal Processing*, 82(12):1903–1915.

Nowotarski, J. and Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568.

Obozinski, G., Taskar, B., and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.

Ochs, P., Chen, Y., Brox, T., and Pock, T. (2014). iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419.

Oliver, H. W. (1954). The exact Peano derivative. *Transactions of the American Mathematical Society*, 76(3):444–456.

Panageas, I. and Piliouras, G. (2016). Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*.

Park, D., Kyrillidis, A., Caramanis, C., and Sanghavi, S. (2016). Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. *arXiv preprint arXiv:1606.03168*.

Park, D. C., El-Sharkawi, M., Marks, R., Atlas, L., and Damborg, M. (1991). Electric load forecasting using an artificial neural network. *IEEE transactions on Power Systems*, 6(2):442–449.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, 1985*, pages 329–334.

Perkins, D. N., Salomon, G., et al. (1992). Transfer of learning. *International encyclopedia of education*, 2:6452–6457.

Pesaran, M. H. and Pick, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics*, 29(2):307–318.

Petra, C. G., Zavala, V., Nino-Ruiz, E., and Anitescu, M. (2014). Economic impacts of wind covariance estimation on power grid operations. *Preprint ANL/MCS-P5M8-0614*.

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, 2011.

Pinson, P. (2012). Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):555–576.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge University Press.

Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653.

Pong, T. K., Tseng, P., Ji, S., and Ye, J. (2010). Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489.

Rai, P., Kumar, A., and Daume, H. (2012). Simultaneously leveraging output and task structures for multiple-output regression. In *Advances in Neural Information Processing Systems*, pages 3185–3193.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Herbert Robbins Selected Papers*, pages 41–47.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

RTE (2011). Référentiel de températures. http://clients.rte-france.com/lang/fr/clients_consommateurs/services/actualites.jsp?id=9482&mode=detail. Last accessed on July 10, 2019.

RTE (2014). Méthodologie des prévisions. http://clients.rte-france.com/lang/fr/visiteurs/vie/courbes_methodologie.jsp. Last accessed on August 29, 2019.

RTE (2016a). L'équilibre offre-demande d'électricité pour l'hiver 2016-2017. Report, RTE.

RTE (2016b). Schéma décennal de développement du réseau 2015. Report, RTE.

RTE (2018). Bilan Électrique. https://bilan-electrique-2018.rte-france.com/production-totale/#. Last accessed on August 29, 2019.

RTE (2019a). Eco2mix. https://rte-france.com/fr/eco2mix/eco2mix-consommation. Last accessed on July 28, 2019.

RTE (2019b). La carte du réseau. https://www.rte-france.com/fr/la-carte-du-reseau. Last accessed on August 28, 2019.

RTE (2019c). L'équilibre offre-demande d'électricité pour l'hiver 2018-2019. Report, RTE.

RTE (2019d). L'équilibre offre-demande d'électricité pour l'été 2019. Report, RTE.

RTE (2019e). RTE en chiffres. https://www.rte-france.com/fr/ecran/1er-reseau-de-transport-d-electricite-d-europe. Last accessed on August 25, 2019.

Sanandaji, B. M., Tascikaraoglu, A., Poolla, K., and Varaiya, P. (2015). Low-dimensional models in spatio-temporal wind speed forecasting. In *2015 American Control Conference (ACC)*, pages 4485–4490. IEEE.

Sangnier, M., Fercoq, O., and d'Alché Buc, F. (2016). Joint quantile regression in vector-valued RKHSs. In *Advances in Neural Information Processing Systems*, pages 3693–3701.

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

She, Y. (2017). Selective factor extraction in high dimensions. *Biometrika*, 104(1):97–110.

Shenoy, S., Gorinevsky, D., and Boyd, S. (2015). Non-parametric regression modeling for stochastic optimization of power grid load forecast. In *2015 American Control Conference (ACC)*, pages 1010–1015. IEEE.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the american statistical association*, 105(489):25–35.

Stewart, G. (2012). Smooth local bases for perturbed eigenspaces. *Institute for Advanced Computer Studies TR*, page 08.

Stone, C. J. and Koo, C.-Y. (1985). Additive splines in statistics. *Proceedings of the American Statistical Association. Original pagination is p*, 45:48.

Sugiyama, M. and Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044.

Sun, J., Qu, Q., and Wright, J. (2015). When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*.

Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, 204(1):139–152.

Taylor, J. W. (2011). Short-term load forecasting with exponentially weighted methods. *IEEE Transactions on Power Systems*, 27(1):458–464.

Thouvenot, V. (2015). *Estimation et sélection pour les modèles additifs et application à la prévision de la consommation électrique*. PhD thesis.

Thouvenot, V., Pichavant, A., Goude, Y., Antoniadis, A., and Poggi, J.-M. (2015). Electricity forecasting using multi-stage estimators of nonlinear additive models. *IEEE Transactions on Power Systems*, 31(5):3665–3673.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567.

Tibshirani, R. J. et al. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Ueda, N. and Nakano, R. (1995). Deterministic annealing variant of the EM algorithm. In *Advances in neural information processing systems*, pages 545–552.

Velu, R. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation theory III*, 2.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Wang, L., Zhang, X., and Gu, Q. (2016). A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*.

Wang, Y., Wipf, D., Ling, Q., Chen, W., and Wassell, I. J. (2015). Multi-task learning for subspace segmentation.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Weron, R. (2007). *Modeling and forecasting electricity loads and prices: A statistical approach*, volume 403. John Wiley & Sons.

Wijaya, T. K. (2015). Pervasive data analytics for sustainable energy systems. Technical report, EPFL.

Wijaya, T. K., Humeau, S. F. R. J., Vasirani, M., and Aberer, K. (2014). Individual, aggregate, and cluster-based aggregate forecasting of residential demand. Technical report, EPFL.

Wijaya, T. K., Sinn, M., and Chen, B. (2015). Forecasting uncertainty in electricity demand. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Wikipedia (2019). Energy demand management. https://en.wikipedia.org/wiki/Energy_demand_management. Last accessed on August 18, 2019.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

Wipf, D. (2014). Non-convex rank minimization via an Empirical Bayesian approach. *arXiv preprint arXiv:1408.2054*.

Wood, S. and Wood, M. S. (2015). Package MGCV. *R package version*, 1:29.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Wytock, M. and Kolter, J. Z. (2013). Large-scale probabilistic forecasting in energy systems using sparse gaussian conditional random fields. In *52nd IEEE Conference on Decision and Control*, pages 1019–1024. IEEE.

Xu, Y. and Yin, W. (2017). A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H., Huang, J., et al. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association*, 96(453):247–259.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. (2017a). The global optimization geometry of low rank matrix optimization. *arXiv preprint arXiv:1703.01256*.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. (2017b). The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 67(2):301–320.