Construction of an informative hierarchical prior distribution. Application to electricity load forecasting

Tristan Launay, Anne Philippe, Sophie Lamarche

Construction of an Informative Hierarchical Prior Distribution: Application to Electricity Load Forecasting

Tristan Launay[1,2] Anne Philippe[1] Sophie Lamarche[2]

[1] Laboratoire de Mathématiques Jean Leray, 2 Rue de la Houssinière – BP 92208, 44322 Nantes Cedex 3, France
[2] Electricité de France R&D, 1 Avenue du Général de Gaulle, 92141 Clamart Cedex, France

Abstract

In this paper, we are interested in the estimation and prediction of a parametric model on a short dataset upon which it is expected to overfit and perform badly. To overcome the lack of data (relatively to the dimension of the model) we propose the construction of a hierarchical informative Bayesian prior based upon another longer dataset which is assumed to share some similarities with the original, short dataset. We apply the methodology to a basic model for the electricity load forecasting on both simulated and real datasets, where it leads to a substantial improvement of the quality of the predictions.

informative prior, hierarchical prior, mcmc algorithms, short dataset, electricity load forecasting

1 Introduction

We are interested in the development of a methodology to improve the estimation and the predictions of a parametric model over a short dataset. The limited size of a dataset coupled with the high dimensionality of a model often leads to a typical overfitting situation : the estimated values are relatively close to the observations while the errors in prediction are an order of magnitude larger or more. This lack of robustness can be somewhat alleviated by the use of a Bayesian estimation relying on an informative prior distribution, but the very fact that the data available is limited makes the posterior distribution all the more sensitive to the choice of that prior.

[1]Laboratoire de Mathématiques Jean Leray, 2 Rue de la Houssinière – BP 92208, 44322 Nantes Cedex 3, France
[2]Electricité de France R&D, 1 Avenue du Général de Gaulle, 92141 Clamart Cedex, France

To design a sensible prior in such a situation, we consider the case where another long dataset is available, upon which the model performs equally well in both estimation and prediction. We assume the long and the short datasets are somehow similar in a non-obvious way. That the similarity between the parameters underlying the two datasets (we will assume they are indeed coming from the model considered) cannot be easily guessed prevents us from trying to model the datasets simultaneously because it would require a rather precise knowledge of the link between the two. We propose a general way of building a hierarchical (see Gelman and Hill, 2007, for a general review on the subject of hierarchical models) informative prior for the short dataset from the long one that goes as follows :

1. we first estimate the posterior distribution on the long dataset using a non-informative prior, arguing that the design of an informative prior for this dataset is not necessary, since the data available is enough to estimate and predict the model in this case ;

2. we extract key informations from this estimation (e.g. moments) to design a hierarchical prior for the short dataset which takes into account the prior information that the datasets are somehow similar, via the introduction of hyperparameters designed to model and estimate this similarity.

As an application we put our method to the test on an unrefined version of a regression model used for the electricity load forecasting in France called EVENTAIL (see Bruhns et al., 2005). Due to the very periodic nature of its regressors, the model typically requires 4 or 5 years of data to provide satisfactory predictions. When the dataset used for the estimation is shorter (think 1 year or less worth of data for the recently started study of a population), we find ourselves in an overfit situation where the prediction errors are way larger than the estimation errors. Although electricity load curves may largely differ from one population to another, they may also share some common features. The latter case is expected to happen when the global population studied is an aggregation of non-homogeneous subpopulations for which the estimations are made harder due to the relative lack of data.

The paper is organised as follows. In Section 2 we focus on the general methodology and describe the way we carried our experimentations, we also present the general regression model used for our tests and applications. In Section 3, we present the semi-conjugated priors (informative and non-informative) used on each of the datasets. The ad hoc MCMC algorithms we developed to estimate the mean and variance of the posterior distributions are push backed into the appendix so as not to obfuscate the main point of the paper by technical details. In Sections 4 and 5, we use these algorithms to illustrate and validate our approach in simulated and real situations : we show the contribution of the informative prior over the precision of both the estimated parameters and the forecasts in the case of a basic electricity load forecasting model.

2 Methodology

2.1 General principle

Let us define here some notations that we shall keep throughout this paper. Hereafter, we denote $\mathcal{B}$ a short dataset over which we would like to estimate the model and we denote $\mathcal{A}$ a long dataset known or thought to share some common features with $\mathcal{B}$. We will denote θ the parameters of the model, $y^{\mathcal{A}}$ the observations from $\mathcal{A}$ and $L(y|\theta)$ the likelihood of the model.

We propose a method designed to help improve parameter estimations and model predictions over $\mathcal{B}$ with the help of $\mathcal{A}$. Let $\pi^{\mathcal{A}}$ be the prior distribution used on $\mathcal{A}$ and $\pi^{\mathcal{A}}(\cdot|y^{\mathcal{A}})$ the associated posterior distribution. Note that the choice of $\pi^{\mathcal{A}}$ is not crucial as long as it remains non-informative enough since the model can be correctly estimated from the data alone on $\mathcal{A}$. We assume that $\pi^{\mathcal{B}}$, the prior distribution to be used on $\mathcal{B}$, is to be chosen within the parametric family

$$\mathcal{F} = \{\pi_\lambda;\ \lambda \in \Lambda\}.$$

Since selecting $\pi^{\mathcal{B}} \in \mathcal{F}$ is equivalent to picking $\lambda^{\mathcal{B}} \in \Lambda$, and since we want $\pi_{\mathcal{B}}$ to retain some key-features of $\pi^{\mathcal{A}}(\cdot|y^{\mathcal{A}})$, we want to pick $\lambda^{\mathcal{B}}$ using some of the information contained inside the posterior distribution obtained on $\mathcal{A}$. We assume that there exists an operator $T : \mathcal{F} \to \Lambda$, such that

$$T[\pi_\lambda] = \lambda,$$

and choose $\lambda^{\mathcal{B}}$ proportional to $T[\pi^{\mathcal{A}}(\cdot|y^{\mathcal{A}})]$, in the sense that

$$\lambda^{\mathcal{B}} = KT[\pi^{\mathcal{A}}(\cdot|y^{\mathcal{A}})],$$

where $K : \Lambda \to \Lambda$ itself is an unknown linear operator that we assume diagonal for ease of use.

The operator K can be interpreted as a similarity operator between $\mathcal{A}$ and $\mathcal{B}$, and its diagonal components as similarity coefficients measuring how close the two datasets really are when looked at through T. The diagonal components of K are hyperparameters of the prior we designed, and we give them a vague hierarchical prior distribution centred around q, the prior on q being vague and centred around 1.

The hyperparameter q may also be regarded as a more global similarity coefficient, since it represents the mean of all the similarity coefficients. The prior mean of q is forced to 1 to reflect the prior knowledge that the datasets are somehow similar. The variance of the prior distribution of q could in theory be reduced, going from a vague prior to a more informative structure, depending on the confidence we have over the similarity between the datasets. We chose not to however, so as to keep the procedure we describe from requiring any delicate subjective adjustments.

We present now two frequent situations where the above procedure can be written in a simpler way.

Example 1 (Method of Moments)**.** *We assume that the elements of $\mathcal{F}$ can be identified via their m first moments : the operator T can then be reduced to a function F of the m first moments operators, i.e. $\lambda = T[\pi_\lambda] = F(\mathbb{E}(\theta), \ldots, \mathbb{E}(\theta^m))$. The expression of $\lambda^{\mathcal{B}}$ then becomes*

$$\lambda^{\mathcal{B}} = KF(\mathbb{E}(\theta|y^{\mathcal{A}}), \ldots, \mathbb{E}(\theta^m|y^{\mathcal{A}})).$$

Note that, if the prior requires the specification of at least the two first moments, even though the priors from the upper layers of the model are vague, the correlation matrix estimated on the dataset $\mathcal{A}$ remains untouched and is directly plugged into in the hierarchical prior if we consider centred moments for orders greater than 1.

Example 2 (Conjugacy)**.** *We consider the case where $\mathcal{F}$ is the family of priors conjugated for the model. If the prior $\pi^{\mathcal{A}}$ belongs to $\mathcal{F}$ then the associated distribution $\pi^{\mathcal{A}}(\cdot|y^{\mathcal{A}})$ does too and there corresponds a parameter $\lambda^{\mathcal{A}}(y^{\mathcal{A}})$ to it. The expression of $\lambda^{\mathcal{B}}$ thus reduces to*

$$\lambda^{\mathcal{B}} = K\lambda^{\mathcal{A}}(y^{\mathcal{A}}).$$

2.2 Description of the model

Modelling and forecasting the electricity load (or demand) on a day-to-day basis has long been a key activity for any company involved in the electricity industry. It is first and foremost needed to supply a fixed voltage at all ends of an electricity grid : to be able to do so, the amount of electricity produced has to match the demand very closely at any given time and experts usually make use of short-term forecasts with this aim in view as mentioned in Cottet and Smith (2003).

Electricity load usually has a large predictable component due to its very strong daily, weekly and yearly periodic behaviour. It has also been noted in many regions that the weather usually affects the load too, the most important meteorological factor typically being the temperature (see Al-Zayer and Al-Ibrahim, 1996, for an example).

The EVENTAIL model (see Bruhns et al., 2005) is a non-linear regression model used to describe and forecast the electricity load in France. For each instant of the day (each instant lasts 30 minutes, starting from 00:00AM), the model that we consider in this paper is made of three components, which we explain briefly in the next paragraphs, and is usually formulated as follows : for

$t = 1, \ldots, N$,

$$
\begin{aligned}
y_t &= x_t^{(1)} x_t^{(2)} + x_t^{(3)} + \epsilon_t \qquad (1)\\
x_t^{(1)} &= \sum_{j=1}^{d_{11}} \left[z_j^{\cos} \cos\left(\frac{2j\pi}{365.25} t\right) + z_j^{\sin} \sin\left(\frac{2j\pi}{365.25} t\right) \right] + \sum_{j=1}^{d_{12}} \omega_j \mathbb{1}_{\Omega_j}(t),\\
x_t^{(2)} &= \sum_{j=1}^{d_2} \psi_j \mathbb{1}_{\Psi_j}(t),\\
x_t^{(3)} &= g(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u),
\end{aligned}
$$

where y_t is the load of day t and where $\epsilon_1, \ldots, \epsilon_N$ are assumed independent and identically distributed with common distribution $\mathcal{N}(0, \sigma^2)$.

The $x^{(1)}$ component is meant to account for the average seasonal behaviour of the electricity load, with a truncated Fourier series (whose coefficients are $z_j^{\cos} \in \mathbb{R}$ and $z_j^{\sin} \in \mathbb{R}$) and gaps (parameters $\omega_j \in \mathbb{R}$) which represent the average levels of electricity load over predetermined periods given by a partition $(\Omega_j)_{j \in \{1,\ldots,d_{12}\}}$ of the calendar. This partition usually specifies holidays, or the period of time when daylight saving time is in effect i.e. major breaks in the electricity consumption behaviour. The left part of Figure 1 shows a typical behaviour over two different periods of time (summer vs. winter).

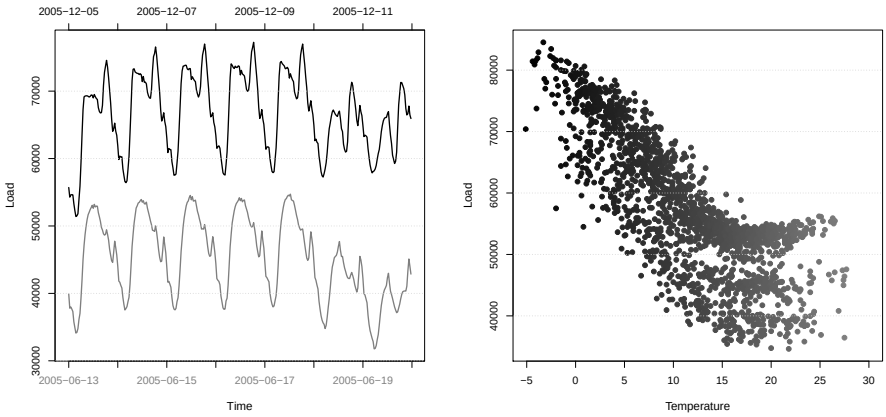

Figure 1: Left : French Electricity load from 13/06/2005 to 29/06/2005 (in grey) and from 05/12/2005 to 11/12/2005 (in black). The load is expressed in MW. Notice the daily patterns of the electricity load are not the same during summer and winter. Right : French Electricity load at 10:00 over 5 years against temperatures. The load seems to increase linearly with the temperature below a certain threshold.

The $x^{(2)}$ component allows for day-to-day adjustments of the seasonal behaviour $x^{(1)}$ through shapes (parameters ψ_j) that depends on the so-called days'

types which are given by a second partition $(\Psi_j)_{j\in\{1,\dots,d_2\}}$ of the calendar. This partition usually separates weekdays from weekends, and bank holidays. The differences between two different daytypes are visible on the left part of Figure 1 too. For obvious identifiability reasons, the vector ψ is restricted to the positive quadrant of the $\|\cdot\|_1$-unit sphere in $\mathbb{R}^{d_2}$, that we denote

$$S_+^{d_2}(0,1) = \left\{\psi \in (\mathbb{R}_+)^{d_2};\ \|\psi\|_1 = 1\right\}.$$

The $x^{(3)}$ component represents the non-linear heating effect that links the electricity load to the temperature (see Seber and Wild, 2003, for a general presentation of non-linear models), with the help of 2 parameters. The heating threshold $u \in [\underline{u}, \overline{u}]$ corresponds to the temperature above which the heating effect is considered null and is usually estimated to be roughly around 15°C. The heating effect is supposed to be linear for temperatures below the threshold and null for temperatures above. The restriction on the support of the threshold u simply expresses the fact that the threshold is sought within the range of the observed temperatures, i.e. $u \in [\underline{u}, \overline{u}]$ with

$$\min_{t=1,\dots,N} T_t < \underline{u} < \overline{u} < \max_{t=1,\dots,N} T_t.$$

The heating gradient $g \in \mathbb{R}^*$ represents the intensity of the heating effect, i.e. the slope (assumed to be non-zero) of the linear part that can be observed on the right part of Figure 1.

The previous model can be re-written in the following condensed and more generic way : for $t = 1, \dots, N$,

$$y_t = (A_{t\bullet}\alpha)(B_{t\bullet}\beta + C_t) + \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u) + \epsilon_t \tag{2}$$

where $\epsilon_1, \dots, \epsilon_N$ are independent and identically distributed with common distribution $\mathcal{N}(0, \sigma^2)$ and where the notation $M_{i\bullet}$ is used to denote the i-row of a matrix M. The matrices A of size $N \times d_A$, B of size $N \times d_\beta$, C of size $N \times 1$, and T of size $N \times 1$ are known exogenous variables while the parameters of the model to be estimated are

$$(\alpha, \beta, \gamma, u, \sigma^2) \in \mathbb{R}^{d_\alpha} \times B_+^{d_\beta}(0,1) \times R^* \times [\underline{u}, \overline{u}] \times \mathbb{R}_+^*,$$

where $B_+^{d_\beta}(0,1) = \{\beta \in (\mathbb{R}_+)^{d_\beta};\ \|\beta\|_1 \leq 1\}$ is the positive quadrant of the $\|\cdot\|_1$-unit ball of dimension d_β.

Remark 3. *Considering this last expression, the model is quite general since the bulk of it could be thought of as the product of two linear regressions, with the added twist of a non-linearity introduced via the threshold parameter u (change-point of the model). Even though the priors and algorithms constructed in the coming sections do depend on the model introduced here, they can be modify in a straightforward manner, should the reader want to tweak the model a bit (e.g. deleting a part or adding a similar one).*

3 Specifications of the priors for the model

3.1 Informative situation

We now present the hierarchical prior we build from $\mathcal{A}$ to improve our predictions on $\mathcal{B}$ for the model at hand. To be able to build it, we assume that we have already collected $\mu^{\mathcal{A}}$ and $\Sigma^{\mathcal{A}}$ the posterior mean and posterior variance of η from a non-informative approach applied to the long dataset $\mathcal{A}$. Hereafter we denote $M^{\mathcal{A}} = \text{diag}(\mu^{\mathcal{A}})$. The hierarchical prior that we propose introduces new parameters to model the similarity between the two datasets. For the sake of clarity, we drop the $^{\mathcal{B}}$ notation : when not explicitly specified, the dataset, data and observations as well as the prior and posterior distributions we refer to in this subsection will be those corresponding to $\mathcal{B}$. We first describe the hierarchical prior we use and then prove that it leads to a proper posterior distribution (see Proposition 4).

Instead of using the obvious and far too rigid prior

$$\eta \sim \mathcal{N}(\mu^{\mathcal{A}}, \Sigma^{\mathcal{A}})$$

we introduce hyperparameters $(k, l) \in \mathbb{R}^d \times \mathbb{R}$ and $(q, r) \in \mathbb{R} \times \mathbb{R}_+^*$ such that

$$\begin{aligned} \eta|k, l &\sim \mathcal{N}(M^{\mathcal{A}}k, l^{-1}\Sigma^{\mathcal{A}}) \\ k|q, r &\sim \mathcal{N}(q(1, \ldots, 1)', r^{-1}I_d) \end{aligned}$$

to allow for more robustness. The coordinates of the vector k can be interpreted as similarity coefficients between parameters of $\mathcal{A}$ and $\mathcal{B}$ and the strictly positive scalar l can be seen as a way to alternatively weaken or strengthen the covariance matrix as needed. Hyperparameters q and r are more general indicators of how close $\mathcal{A}$ and $\mathcal{B}$ are, q corresponding to the mean of the coordinates of k and r being their inverse-variance. l, q, r and σ^2 of course require a prior distribution too. For σ^2 we use a non-informative prior (we chose $\pi(\sigma^2) = \sigma^{-2}$) because we do not want to make any kind of assumptions about the noise around both datasets. This prior is non-informative in the sense that it matches Jeffreys' prior distribution on σ^2 for a Gaussian linear regression. For the three other parameters, based on semi-conjugacy considerations, we use :

$$l \sim \mathcal{G}(a_l, b_l), \qquad q \sim \mathcal{N}(1, \sigma_q^2), \qquad r \sim \mathcal{G}(a_r, b_r), \tag{3}$$

where a_l, b_l, a_r, b_r and σ_q^2 are fixed positive real numbers such that the prior distribution on l, q and r are vague. These prior distributions are chosen because of their conjugacy properties (as will be seen in the MCMC algorithm). The vagueness requirement that we impose on these priors is motivated by the fact that we want to keep as general a framework as possible without having to tweak each and every prior coefficient for different applications.

The hierarchical prior that we use is built as follows :

$$\pi(\theta, k, l, q, r) \propto \pi(\eta|k, l)\pi(k|q, r)\pi(l)\pi(q)\pi(r)\pi(\sigma^2) \tag{4}$$

with

$$\begin{aligned}
\pi(\sigma^2) &\propto \sigma^{-2} \\
\pi(\eta|k,l) &\propto l^{\frac{d}{2}} \exp\left(-\frac{1}{2}(\theta - M^{\mathcal{A}}k)' l (\Sigma^{\mathcal{A}})^{-1}(\theta - M^{\mathcal{A}}k)\right) \\
\pi(k|q,r) &\propto |r|^{\frac{d}{2}} \exp\left(-\frac{1}{2} r \sum_{i=1}^{d} (k_i - q)^2\right) \\
\pi(l) &\propto l^{a_l - 1} \exp\left(-b_l l\right) \mathbb{1}_{\mathbb{R}_+^*}(l) \\
\pi(q) &\propto |\sigma_q^{-2}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sigma_q^{-2}(q-1)^2\right) \\
\pi(r) &\propto r^{a_r - 1} \exp\left(-b_r r\right) \mathbb{1}_{\mathbb{R}_+^*}(r).
\end{aligned}$$

The posterior measure is hence given by

$$\begin{aligned}
\pi(\theta, k, l, q, r|y, \mathcal{D}) &\propto f(y|\theta, \mathcal{D})\pi(\theta, k, l, q, r) \qquad (5) \\
&\propto \sigma^{-N-2} \exp\left(-\frac{1}{2}\sigma^{-2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right) \mathbb{1}_{[0,\,1]\times[\underline{u},\,\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1, u, \sigma^2) \\
&\quad \times |r|^{\frac{d}{2}} \exp\left(-\frac{1}{2} r \sum_{i=1}^{d} (k_i - q)^2\right) l^{a_l - 1} \exp\left(-b_l l\right) \mathbb{1}_{\mathbb{R}_+^*}(l) \\
&\quad \times |\sigma_q^{-2}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sigma_q^{-2}(q-1)^2\right) r^{a_r - 1} \exp\left(-b_r r\right) \mathbb{1}_{\mathbb{R}_+^*}(r).
\end{aligned}$$

Proposition 4. *For $(\beta, u) \in B_+^{d_\beta}(0,1)\times[\underline{u},\,\overline{u}]$ denote $A_*(\beta, u)$ the matrix whose rows are*

$$(A_*)_{t\bullet}(\beta, u) = \left[(B_{t\bullet}\beta + C_t)A_{t\bullet}, (T_t - u)\mathbb{1}_{[T_t,\,+\infty[}(u)\right], \quad t = 1, \ldots, N,$$

and suppose $A_*'(b,u)A_*(b,u)$ has full rank for every $(\beta, u) \in B_+^{d_\beta}(0,1) \times [\underline{u},\,\overline{u}]$. Assume furthermore that $N > d_\alpha + 1$ and that $(y_1, \ldots, y_N)$ are observations coming from the model (2) and the posterior measure (5) is then a well-defined (proper) probability distribution.

Proof. First notice that $\int \pi(\theta, k, l, q, r|y, \mathcal{D})\,\mathrm{d}\sigma^2$ is proportional to

$$\|y - \mu(\eta|\mathcal{D})\|_2^{-N} \mathbb{1}_{[0,\,1]}(\|\beta\|_1)\mathbb{1}_{[\underline{u},\,\overline{u}]}(u)\pi(\eta|k,l)\pi(k|q,r)\pi(l)\pi(q)\pi(r),$$

for almost every y and that the function $\theta \mapsto \|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ is bounded, for almost every y. The posterior integrability is hence trivial as long as $\pi(\eta|k,l)\pi(k|q,r)\pi(l)\pi(q)\pi(r)$ itself is a proper distribution which is the case here. □

3.2 Non-informative situation

We propose here a non-informative prior to use with the long dataset $\mathcal{A}$. Note that since the dataset $\mathcal{A}$ is long enough, the choice of the prior distribution

used in this situation does not matter much as long as it remains vague enough. For the sake of clarity again, we drop the $^{\mathcal{A}}$ notation : when not explicitly specified, the dataset, data and observations as well as the prior and posterior distributions we refer to in this subsection will be those corresponding to $\mathcal{A}$. We show that the use of a non-informative prior distribution leads to a proper posterior distribution (see Proposition 5).

We use the following non-informative prior

$$\pi(\theta) \propto \sigma^{-2}.$$

This prior is non-informative in the sense that it matches Jeffreys' prior distribution on σ^2 for a Gaussian linear regression and matches Laplace's flat prior on the other parameters. It leads to the following posterior distribution

$$\begin{aligned}\pi(\theta|y,\mathcal{D}) &\propto L(y|\theta,\mathcal{D})\pi(\theta) \\ &\propto \sigma^{-N-2}\exp\left(-\frac{1}{2}\sigma^{-2}\|y-\mu(\eta|\mathcal{D})\|_2^2\right)\mathbb{1}_{[0,\,1]\times[\underline{u},\,\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1,u,\sigma^2).\end{aligned} \tag{6}$$

Proposition 5. *For $(\beta,u)\in B_+^{d_\beta}(0,1)\times[\underline{u},\,\overline{u}]$ denote $A_*(\beta,u)$ the matrix whose rows are*

$$(A_*)_{t\bullet}(\beta,u) = \left[(B_{t\bullet}\beta+C_t)A_{t\bullet}, (T_t-u)\mathbb{1}_{[T_t,\,+\infty[}(u)\right], \quad t=1,\dots,N,$$

and suppose $A_*'(b,u)A_*(b,u)$ has full rank for every $(\beta,u)\in B_+^{d_\beta}(0,1)\times[\underline{u},\,\overline{u}]$. Assume furthermore that $N > d_\alpha+1$ and that $(y_1,\dots,y_N)$ are observations coming from the model (2), the posterior measure (6) is then a well-defined (proper) probability distribution.

Proof. Notice first that

$$\int \pi(\eta,\sigma^2|y,\mathcal{D})\,\mathrm{d}\sigma^2 \propto \|y-\mu(\eta|\mathcal{D})\|_2^{-N}\mathbb{1}_{[0,\,1]}(\|b\|_1)\mathbb{1}_{[\underline{u},\,\overline{u}]}(u) \quad \text{for almost every } y,$$

and observe then that

$$\|y-\mu(\eta|\mathcal{D})\|_2^2 = \sum_{t=1}^N \left[y_t-(B_{t\bullet}\beta+C_t)A_{t\bullet}\alpha-(T_t-u)\mathbb{1}_{[T_t,\,+\infty[}(u)\gamma\right]^2.$$

Let $(\beta_0,u_0)\in B_+^{d_\beta}(0,1)\times[\underline{u},\,\overline{u}]$ and denote $\alpha_*=(\alpha,\gamma)$. We write

$$\begin{aligned}\|y-\mu((\alpha,\beta_0,\gamma,u_0)|\mathcal{D})\|_2^2 &= \sum_{t=1}^N \left[y_t-(B_{t\bullet}\beta_0+C_t)A_{t\bullet}\alpha-(T_t-u_0)\mathbb{1}_{[T_t,\,+\infty[}(u_0)\gamma\right]^2 \\ &= \|y-A_*(\beta_0,u_0)\alpha_*\|_2^2,\end{aligned}$$

and thus obtain the following equivalence, as $(\beta,u)\to(\beta_0,u_0)$ and $\|\alpha_*\|_2\to+\infty$

$$\|y-\mu(\eta|\mathcal{D})\|_2^{-N} \sim \|y-A_*(\beta_0,u_0)\alpha_*\|_2^{-N}. \tag{7}$$

The triangular inequality applied to the right hand side of (7) gives

$$\|y - A_*(\beta_0, u_0)\alpha_*\|_2^{-N} \leq \big|\|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2\big|^{-N}. \tag{8}$$

Since $A'_*(\beta_0, u_0)A_*(\beta_0, u_0)$ has full rank, by straightforward algebra we get

$$\lambda\|\alpha_*\|_2^2 \leq \|A_*(\beta_0, u_0)\alpha_*\|_2^2,$$

where λ is the smallest eigenvalue $(A_*(\beta_0, u_0))'A_*(\beta_0, u_0)$ and is strictly positive. We can hence find an equivalent of the right hand side of (8) as $\|\alpha_*\|_2 \to +\infty$, which is

$$\big|\|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2\big|^{-N} \sim \lambda^{-N/2}\|\alpha_*\|_2^{-N}. \tag{9}$$

Combining (7), (8) and (9) together, we see that the integrability of the left hand side of (7) as $(\beta, u) \to (\beta_0, u_0)$ and $\|\alpha_*\|_2 \to +\infty$ is directly implied by that of $\|\alpha_*\|_2^{-N}$. The latter is of course immediate for $N > d_\alpha + 1$ as can be seen via a quick cartesian to hyperspherical re-parametrisation.

The previous paragraph thus ensures the integrability of $\|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ over sets of the form

$$\{(\beta, u) \in V((\beta_0, u_0)), \|\alpha_*\|_2 \in]M(\beta_0, u_0), +\infty[\}, \quad \forall(\beta_0, u_0) \in B_+^{d_\beta}(0,1) \times [\underline{u}, \overline{u}]$$

where the subset $V((b_0, u_0))$ is an open neighbourhood of (β_0, u_0) and $M(\beta_0, u_0)$ is a real number depending on (β_0, u_0). By compacity of $B_+^{d_\beta}(0,1) \times [\underline{u}, \overline{u}]$ there exists a finite union of such $V((\beta_i, u_i))$ that covers $B_+^{d_\beta}(0,1) \times [\underline{u}, \overline{u}]$. Denoting M the maximum of $M(\beta_i, u_i)$ over the corresponding finite subset of (β_i, u_i), we finally obtain the integrability of $\|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ over $\{(\beta, u) \in B_+^{d_\beta}(0,1), \|\alpha_*\| \in]M, +\infty[\}$.

The integrability of $\|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ over $\{(\beta, u) \in B_+^{d_\beta}(0,1), \|\alpha_*\| \in [0, M]\}$ is trivial, recalling that $\eta \mapsto \|y - \mu(\eta|\mathcal{D})\|_2$ is continuous and does not vanish over this compact for almost every y, meaning its inverse shares these same properties. □

Remark 6. *The condition "A'_*A_* has full rank" mentioned above is typically verified in our applications for the regressors used in the* EVENTAIL *model. To see this, call "vector of heating degrees" the vector whose coordinates are $(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u)$, then not verifying the aforementionned condition is equivalent to saying that "there exists an index i and a threshold u such that the family of vectors formed by the regressors A and the vector of heating degrees is linearly dependant over the subset Ψ_i of the calendar".*

4 Numerical evaluations of the performance on simulated data

For any estimation (posterior mean and variance) on a dataset (be it $\mathcal{A}$ or $\mathcal{B}$), the MCMC algorithms would typically run for 500,000 iterations after a small burn-in period.

4.1 Comparing the hierarchical and the non-informative approaches

Predictive distribution. The Bayesian framework allows us to compute so-called predictive distributions, i.e. the distributions of future observations given past observations. Given a prior distribution $\pi(\theta)$ and the corresponding posterior distribution $\pi(\theta|y,\mathcal{D})$ related to the past observations $y=(y_1,\ldots,y_N)$ and data $\mathcal{D}=[\mathcal{D}_1,\ldots,\mathcal{D}_N]$, the predictive distribution for the future observation y_{N+k}, given data $\mathcal{D}_{N+k}$ is defined as

$$g(y_{N+k}|\mathcal{D}_{N+k},y,\mathcal{D}) := \int f(y_{N+k}|\theta,\mathcal{D}_{N+k})\pi(\theta|y,\mathcal{D})\,\mathrm{d}\theta,$$

and the optimal prediction for the L^2 risk is then :

$$\widehat{y}_{N+k} := \mathbb{E}^{\pi}[y_{N+k}|\mathcal{D}_{N+k},y,\mathcal{D}] \tag{10}$$

$$= \int y_{N+k}g(y_{N+k}|\mathcal{D}_{N+k},y,\mathcal{D})\,\mathrm{d}y_{N+k}. \tag{11}$$

The comparison criterion. To assess the quality of the estimation of the model with our hierarchical prior with regard to the estimation of the model with the non-informative prior, we compare both results based on the quality of the predictions. Let y_{N+1} be the next upcoming observation, corresponding to data $\mathcal{D}_{N+1}$ and observe now that the prediction error can be written as

$$y_{N+1}-\widehat{y}_{N+1} = [y_{N+1}-\mu(\eta_0|\mathcal{D}_{N+1})] + [\mu(\eta_0|\mathcal{D}_{N+1})-\widehat{y}_{N+1}],$$

which expresses the prediction error as a sum of a noise $y_{N+1}-\mu(\eta_0|\mathcal{D}_{N+1})$ (whose theoretical distribution is $\mathcal{N}(0,\sigma^2)$) and a bias which can be seen as an estimation error over the prediction $\mu(\eta_0|\mathcal{D}_{N+1})-\widehat{y}_{N+1}$. We focus solely on the second part, since the first part (the noise) is unavoidable in real situation. Given that we want to validate our model on simulated data, the quantity $\mu(\eta_0|\mathcal{D}_{N+1})-\widehat{y}_{N+1}$ is indeed accessible here whereas it would not be in real situation.

We thus choose to consider the quadratic distance between the real and the predicted model over a year as our quality criterion for a model, i.e. :

$$\sqrt{\frac{1}{365}\sum_{i=1}^{365}[\mu(\eta_0|\mathcal{D}_{N+i})-\widehat{y}_{N+i}]^2}. \tag{12}$$

4.2 Construction of simulated datasets

Both datasets $\mathcal{A}$ and $\mathcal{B}$ were simulated according to the model (1) given on page 5 with $d_{11}=4$ (4 frequencies used for the truncated Fourier series). The calendars and the partitions used for $\mathcal{A}$ and $\mathcal{B}$ were designed to include 7 daytypes ($d_2=7$, one daytype for each day of the week), but did not include any special

days such as bankholidays. They also included 2 offsets ($d_{12} = 2$) to simulate the daylight saving time effect. In the end we thus had $d_\alpha = 4 \times 2 + 2 = 10$ and $d_\beta = 6$ i.e. $d = 19$ using the expression of the model given in (2).

Dataset $\mathcal{A}$. We simulated 4 years of daily data for $\mathcal{A}$ with parameters :

$$\begin{aligned} \sigma^{\mathcal{A}} &= 2, \\ \text{seasonal}: \alpha^{\mathcal{A}} &= (27, 7, -3, 1, 5, -1, 4, 0.5, 490, 495), \\ \text{shape}: \beta^{\mathcal{A}} &= (0.13, 0.15, 0.16, 0.16, 0.16, 0.13), \\ \text{heating}: \gamma^{\mathcal{A}} &= -3, \\ u^{\mathcal{A}} &= 14. \end{aligned}$$

These values were chosen to approximately mimic the typical electricity load of France up to a scaling factor. The temperatures we used for the estimation over $\mathcal{A}$ are those measured from September 1996 to August 2000 at 10:00AM.

Dataset $\mathcal{B}$. We simulated 1 year of daily data for $\mathcal{B}$ with parameters :

$$\begin{aligned} \sigma^{\mathcal{B}} &= 2, \\ \text{seasonal}: \alpha_i^{\mathcal{B}} &= k_\alpha \alpha_i^{\mathcal{A}}, && && \forall i = 1, \ldots, d_\alpha \\ \text{shape}: \beta_1^{\mathcal{B}} &= k_\beta \beta_1^{\mathcal{A}}, && \beta_j^{\mathcal{B}} = \beta_j^{\mathcal{A}}, && \forall j = 2, \ldots, d_\beta \\ \text{heating}: \gamma^{\mathcal{B}} &= k_\gamma \gamma^{\mathcal{A}}, \\ u^{\mathcal{B}} &= k_u u^{\mathcal{A}}. \end{aligned}$$

where the coordinates of the true hyperparameters k were allowed to vary around 1. The temperatures we used for the estimation over $\mathcal{B}$ are those measured from September 2000 to August 2001 at 10:00AM.

We also simulated an extra year of daily data $\mathcal{B}$ for prediction, with the same parameters but using the so-called normal temperatures, meaning that for each day of this extra year the temperature is the mean of all the past temperatures at the same time of the year. We made such a choice to try and suppress any dependency between our simulated results and the chosen temperature for this fictive year of prediction, since we did not want to bias our results because of a rigorous winter or an excessively hot summer.

4.3 Results

We chose to use vague priors (i.e. proper distributions with large variances) for the uppermost layers of our hierarchical prior, and thus decided to use the values :

$$\sigma_q = 10^2, \qquad a_r = b_r = 10^{-6}, \qquad a_l = b_l = 10^{-3}.$$

A study of the Bayesian hierarchical model's sensitivity to these values showed that changing these hyperparameters to achieve prior variances of greater magnitudes hardly influenced the posterior results (means and variances) at all.

This is why we decided to stick to these values for the remainder of our experimentations.

Estimation. We benchmarked the Bayesian model with its hierarchical prior against its original non-informative prior counterpart for different choices of true hyperparameters k over 300 replications (data being simulated anew for each replication), i.e. we simulated many different datasets $\mathcal{B}$ looking more or less similar to $\mathcal{A}$ and applied our method on them. Figure 2 shows the posterior error of η (posterior mean minus the true value) of η, based on 300 replications that correspond to the case where $k_\alpha = k_\beta = k_\gamma = k_u = 1$ i.e $\eta_{\mathcal{A}} = \eta_{\mathcal{B}}$ for both the informative (leftmost) and non-informative (rightmost) method. Marginal confidence interval for the posterior means are much smaller when using the hierarchical prior (most of them hitting the true value). The marginal posterior standard deviations (not shown here) are also reduced when the informative hierarchical prior is used instead of the non-informative prior.

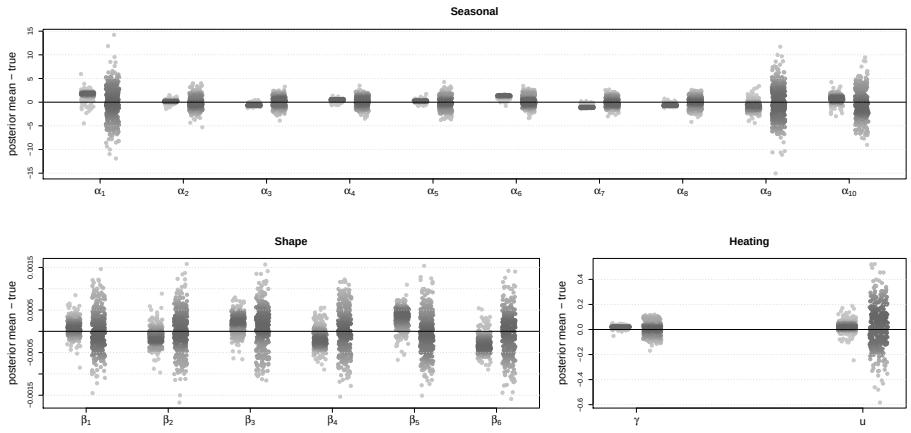

Figure 2: The posterior error (posterior mean minus true value) of α (seasonal parameters), β (shape parameters), and γ and u (heating parameters), based on 300 replications. Leftmost replications correspond to the hierarchical method while the rightmost replications correspond to the non-informative method. Here $k_\alpha = k_\beta = k_\gamma = k_u = 1$.

When the situation is far from being as ideal as the one mentioned above, the hierarchical approach still shows improvement over the non-informative approach but to a lesser extent. Figure 3 shows that the estimations of some of the parameters of the model are improved with the addition of the prior information (α and u) while some are not (β and γ) in the case where $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$. Situations such as $k_\alpha = k_\gamma = k_u = 1$ and $k_\beta = 0.5$ or $k_\alpha = k_\gamma = k_\beta = 1$ and $k_u = 0.5$ were studied too and yielded very similar results i.e. lesser improvements on the estimations of some parameters only. Note that when some coordinates of k are valued to 0.5 while some are valued to 1,

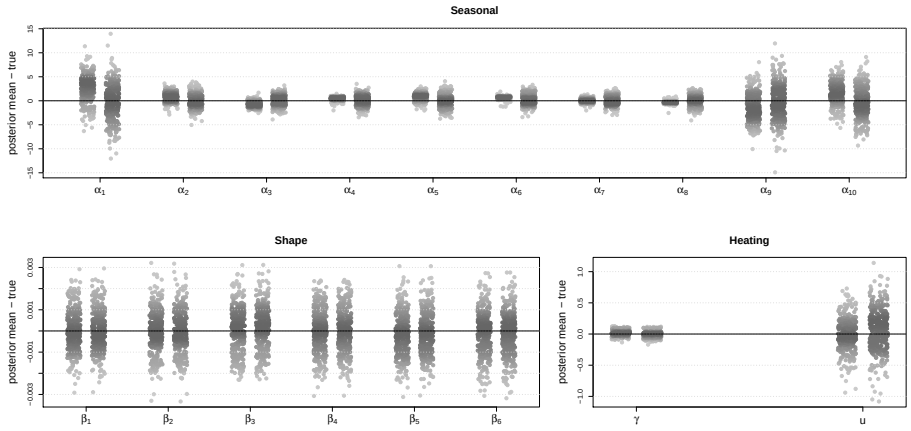

Figure 3: Same caption as in Figure 2 except $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$.

the "similarity" between $\mathcal{A}$ and $\mathcal{B}$ is very weak. The strength or weakness of the similarity between $\mathcal{A}$ and $\mathcal{B}$ cannot be diagnosed directly from the posterior mean of k itself but we will see that the estimations of the hyperparameters q and r may provide a partial answer to this question.

We also estimated the hyperparameters (see Section 3.1 for the specifications of k, l, r) when the hierarchical prior was used. Let us first study the hyperparameter k. Its coordinates seem correctly estimated for the ideal situation where $k_\alpha = k_\beta = k_\gamma = k_u = 1$ as illustrated in the top row of Figure 4 which shows the posterior error of k. When $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$, the estimations obtained are of lesser quality as demonstrated in the bottom row of Figure 4 : most of the seasonal similarity coefficients appear to be biased (while the posterior standard deviation on each coordinate, not shown here, are greater than in the ideal situation). These estimations may thus be used to quantify the closeness of the two datasets.

The estimation of the hyperparameter l itself does not seem to provide a lot of information about the data : during our simulations, its mean value exhibited a lot of variability around the same value over the 300 replications for each of the five simulated scenarios and no reasonable conclusion could be drawn from it.

On the other hand, the estimation of the hyperparameter q does reveal a bit of information about the two datasets $\mathcal{A}$ and $\mathcal{B}$. It is the mean of the coordinates of k on the real axis, as can be seen in the definition of the hierarchical prior in (4) on page 7. However its use remains somewhat limited in the sense that the parameters β of the two datasets are most often very close (meaning the coordinates of k that correspond to them is likely close to 1) while other parameters may vary greatly. Hence even though q provides information about the similarity between $\mathcal{A}$ and $\mathcal{B}$, it cannot be interpreted alone and has to be

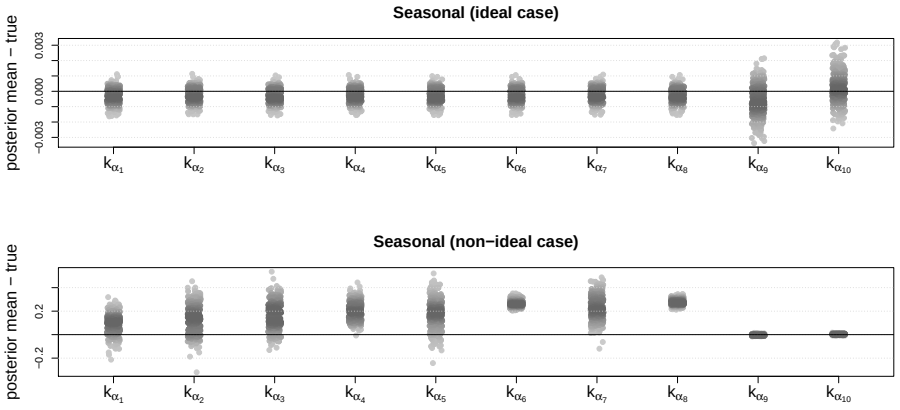

Figure 4: The posterior error (posterior mean minus true value) of k_α (seasonal parameters), based on 300 replications. Top row is for the case were $k_\alpha = k_\beta = k_\gamma = k_u = 1$ and bottom row is for the case where $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$. Leftmost replications correspond to the hierarchical method while the rightmost replications correspond to the non-informative method. Posterior errors of k_β (shape parameters), and k_γ and k_u (heating parameters) are not shown here because no significant deviation from 0 was found on either of these coordinates when the informative prior was used in either case (the empirical variances on these coordinates were bigger in the non-ideal case though, in a similar fashion to what we observe here for k_α).

considered jointly with r. The left part of Figure 5 shows the evolution of the posterior mean of q as $k_\alpha = k_u$ ranges over [0.5, 1].

The estimation of the hyperparameter r (inverse-variance of the prior distribution on k, see (4) again) does in fact reveal some information about the two datasets too. It is a measure of dispersion of k around q, in the sense that the (higher it is, the closer to q the coordinates of k should be. Just like q is the mean of the coordinates of k, r is in fact their inverse-variance. The right part of Figure 5 shows a clear decline of r when $k_\alpha = k_u$ moves away from the ideal value 1 i.e. when the similarity between the datasets $\mathcal{A}$ and $\mathcal{B}$ decrease from strong to weak.

As we previously stated, the similarity between the two datasets has to be assessed simultaneously with q and r and not q only : the mean q could be close to 1, possibly hinting at a perfect similarity between the two datasets, while the variance $1/r$ could be great which would then indicate huge differences between the two estimated sets of parameters for the two datasets.

Prediction. We compared the hierarchical and the non-informative models using our comparison criterion defined in (12) and computing the ratio betwen the two models for different values of k_α and k_γ, k_β and k_u being both set

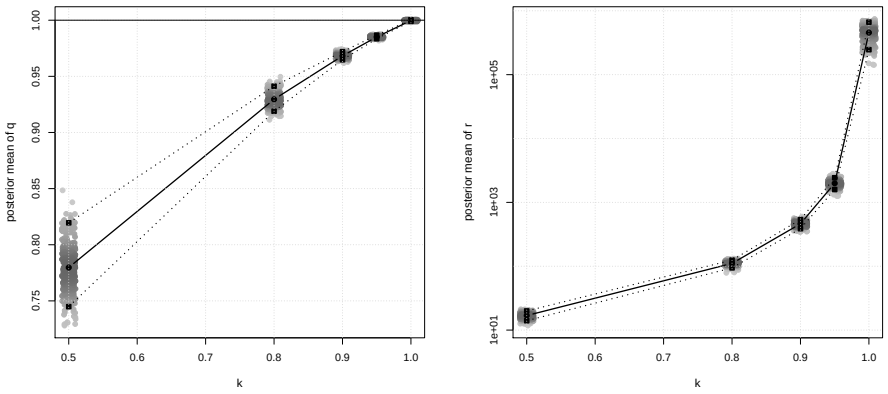

Figure 5: In grey : posterior mean of q (left) and r (right, on a log scale) for the hierarchical prior (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $k_\alpha = k_\gamma$ tested. In black : the circles correspond to the averages, while the squares correspond to the 5% and 95% empirical quantiles. Here $k_\beta = k_u = 1$.

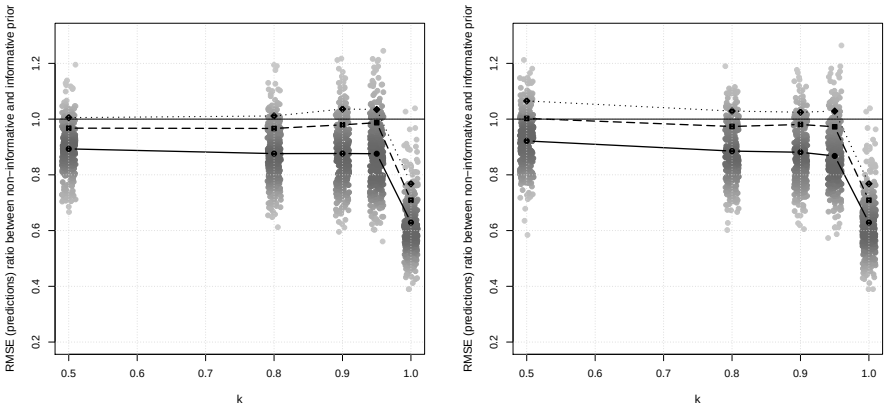

Figure 6: In grey : ratio between error predictions for the hierarchical and the non-informative approach (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $k_\alpha = k_\gamma$ (left, where $k_\beta = k_u = 1$) and k_u (right, where $k_\alpha = k_\beta = k_\gamma = 1$) tested. In black : circles correspond to the averages, while squares and diamonds correspond to the 80% and 90% empirical quantiles of these ratios.

to 1. The left part of Figure 6 shows the results we obtained for k_α and k_γ simultaneously set to the values $1, 0.95, 0.90, 0.80$ and 0.50. Note that since the results appeared to be approximately symmetric with regard to 1 (i.e. for values $1, 1.05, 1.10, 1.20$ and 1.50), we only include one side of the graph in the present article.

On average, the Bayesian hierarchical model is a clear improvement over the Bayesian non-informative one, its performances being maximised when the parameters $\eta^{\mathcal{A}}$ and $\eta^{\mathcal{B}}$ are identical (which is the ideal situation). The performances in prediction are obviously somewhat weakened when the difference between the parameters $\eta^{\mathcal{A}}$ and $\eta^{\mathcal{B}}$ grows greater, but the use of the hierarchical model still leads to an average improvement of 15% over the non-informative model, as can be seen on Figure 6. The results obtained when k_β or k_u are varying while the other coordinates of k are fixed to 1 were very similar (see for example the right part of Figure 6).

5 Application

The dataset we used for $\mathcal{A}$ corresponds to a specific population in France frequently referred to as "non-metered" because their electricity consumption is not directly observed but instead derived as the difference between the overall electricity consumption and the consumption of the "metered" population. We tested our method on two different populations $\mathcal{B} : \mathcal{B}_1$ which is a subpopulation of $\mathcal{A}$ and $\mathcal{B}_2$ which roughly covers the same people that $\mathcal{A}$ does. The sizes (in days) of the datasets are given in the Table 1 below.

To use the model on the datasets, we kept only one load value per day (the results shown hereafter were obtained for the load at 10:00AM). One could, without difficulty, add a cooling effect (symmetric to the heating effect) to the model. We did not consider such an addition here, since the cooling effect remains far less important than the heating effect in France, at the present moment.

$\mathcal{A}$	$\mathcal{B}_1$	$\mathcal{B}_2$
833	207	151

Table 1: Sizes of the real datasets (in days).

We kept the last 30 days of each $\mathcal{B}$ out of the estimation datasets and assessed the model quality over the predictions for those 30 days. It might seem an arbitrary choice and it is indeed, but the important lack of data prevented us from keeping 365 days as we previously did during the simulations. The procedure is similar in spirit to that developped for the simulations, but the results obtained in this section might be dependant on the temperature of these days, or their position in the calendar, while we did our best to avoid such a thing in the simulations. Restricting the prediction period to such a tiny time window might thus weaken somewhat the robustness of our method, but we

nonetheless decided to show the performances we obtained on the real data in this paper.

5.1 Results on the long dataset $\mathcal{A}$

Using the non-informative prior over dataset $\mathcal{A}$ we are able to retrieve estimated predictive densities for future observations or alternatively we can estimate the quantiles of each of these densities to define credibility regions around the predictive mean. Most of the true observations lay well within the boundaries of the 95% credibility intervals of the predictions as can be seen on Figure 7.

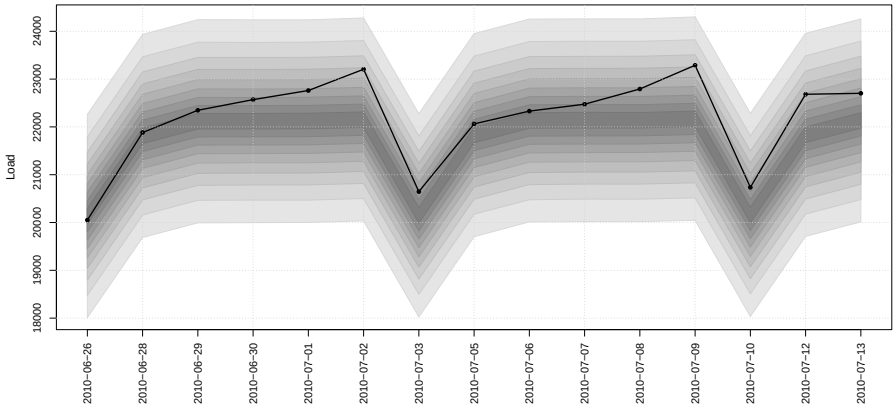

Figure 7: Estimated credibility regions for predictive densities for a few future days of the dataset $\mathcal{A}$. Future observed values are linked together with a black-line. The quantiles are drawn and linked in increasing shades of grey from 45% to 5% and from 55% to 95%.

5.2 Results on the short datasets $\mathcal{B}$

The estimation and prediction errors we obtained for the non-informative and the hierarchical methods on the two datasets $\mathcal{B}$ considered here are given in Table 2 below. While slightly degrading the quality of the fit on the estimation part compared to the non-informative approach, the hierarchical method vastly improves the quality of the predictions, reaching over 50% reduction for the root mean square error (RMSE) measure of accuracy.

The hierarchical prior allows us to retrieve information about the similarity between datasets $\mathcal{A}$ and $\mathcal{B}$ via the estimation of the posterior densities of the hyperparameters. The estimations of the posterior marginal distributions of k are presented on Figure 8 for both $\mathcal{B}_1$ and $\mathcal{B}_2$ and show how these datasets differ.

$\mathcal{B}_1$	non-informative	hierarchical	comparison
RMSE est.	775.93	786.97	+1.42%
RMSE pred.	1863.25	894.00	−52.01%
MAPE est.	4.00	3.93	−0.07
MAPE pred.	19.37	9.30	−10.07

$\mathcal{B}_2$	non-informative	hierarchical	comparison
RMSE est.	1127.60	1202.32	+6.62%
RMSE pred.	2286.42	1339.14	−41.83%
MAPE est.	2.82	2.98	+0.15
MAPE pred.	8.65	3.48	−5.17

Table 2: Results for the dataset $\mathcal{B}_1$ (top) and $\mathcal{B}_2$ (bottom). RMSE is the "root mean square error" and MAPE is the "mean absolute percentage error". Both of these common measures of accuracy were computed on the estimation (est.) and prediction (pred.) parts of the two datasets.

While the coordinates of k related to β seems to lie around 1, the rest of these coordinates do not concentrate around 1 for the dataset $\mathcal{B}_1$ as can be seen on the figures provided : the gaps ω_j of the model EVENTAIL defined in (1) are clearly centred around 0.5 while the rest of the coefficients linger somewhere around 0.7 or 0.8. Unlike $\mathcal{B}_1$, it seems $\mathcal{B}_2$ shares a lot of common features with $\mathcal{A}$: each marginal posterior density of k is peaked around 1 for $\mathcal{B}_2$ which indicates strong similarities. It is possible to derive credibility intervals on the mean values for each coordinate of k and these intervals are found to be smaller on $\mathcal{B}_2$ than they are on $\mathcal{B}_1$, as attested by the sharpness of the densities which are much more peaked on the former dataset than they are on the latter.

The same conclusion can be drawn from the Table 3 in which we listed the estimated posterior means of l, q and r for $\mathcal{B}_1$ and $\mathcal{B}_2$: the estimated value of q (mean of all the coordinates of k) is closer to 1 on the second dataset than on the first and the estimated value of r (inverse-variance of all the coordinates of k) is greater too. These two hyperparameters can thus be used to quickly assess the strength of the similarity between the two datasets $\mathcal{A}$ and $\mathcal{B}$ while only a close study of the posterior marginal densities of k can reveal which coordinates are similar and which are not.

In fact the upper row of Figure 8 suggests that the specification of the hierarchical prior as a mixture of normal distributions $\mathcal{N}(q_i, r_i)$ could possibly help in getting even better results on dataset $\mathcal{B}_1$, to help distinguish at least two groups for the coordinates of k using their means : the coordinates that are close to 1, and those that are not.

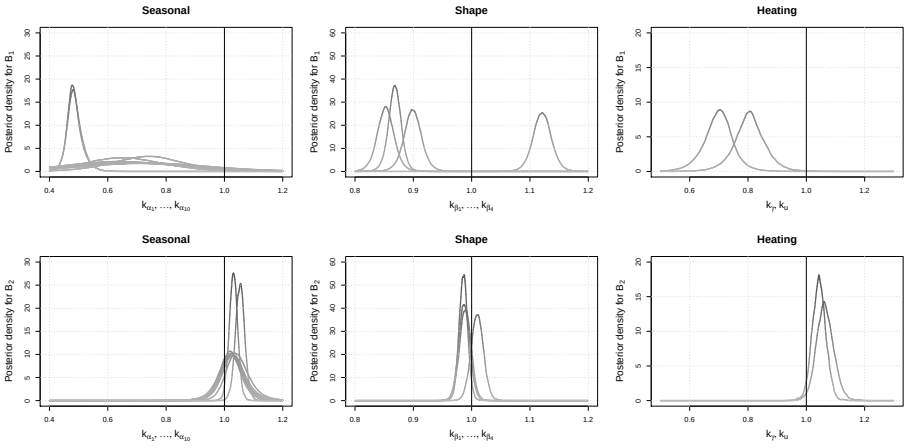

Figure 8: Estimated posterior marginal distributions of k for the hierarchical method and for both datasets $\mathcal{B}_1$ (upper row) and $\mathcal{B}_2$ (lower row). Coefficients corresponding to α, β and (γ, u) are shown on separate graphs.

l	19.17	128.60
q	0.73	1.02
r	24.48	795.16

Table 3: Estimated posterior mean of the hyperparameters l, q and r for both of the studied datasets. These estimations may serve as a summary of the studies : the similarity between $\mathcal{A}$ and $\mathcal{B}_2$ is found to be stronger than the one between $\mathcal{A}$ and $\mathcal{B}_1$ as the posterior mean of q (mean of the similarity coefficients k_i) and r (inverse-variance of the similarity coefficients k_i) indicate together.

6 Appendix

The two MCMC algorithms presented below were developed because direct simulations from the posterior distribution were not possible. The justifications are given after the algorithms themselves. Notice that the full conditional distributions of all the parameters but the threshold u appear to be common distributions in both cases, due to the presence of multiple semi-conjugacy situations. We used a Metropolis-within-Gibbs algorithm (see Marin and Robert, 2007, page 96, for a quick description) based on Gibbs sampling steps for every parameter but u for which we use a Metropolis-Hasting step based on a gaussian random walk proposal.

6.1 Technical Lemmas

Definition 7 (Gaussian conjugacy operator)**.** *We define the (commutative and associative) operator $*$ as*

$$\begin{pmatrix} \mu_1 \\ \Sigma_1 \end{pmatrix} * \begin{pmatrix} \mu_2 \\ \Sigma_2 \end{pmatrix} = \begin{pmatrix} [\Sigma_1^{-1} + \Sigma_2^{-1}]^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) \\ [\Sigma_1^{-1} + \Sigma_2^{-1}]^{-1} \end{pmatrix}$$

for any vectors μ_1 and μ_2 in $\mathbb{R}^d$, for any symmetric positive definite matrices Σ_1 and Σ_2 of size $d \times d$.

Lemma 8 (Conjugacy)**.** *Let X_1 and X_2 be two random truncated Gaussian vectors in $\mathbb{R}^d$*

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_1, S_1)$$
$$X_2 \sim \mathcal{N}(\mu_2, \Sigma_2, S_2)$$

and denote f_1 and f_2 their respective densities, then $f_1 f_2$ is integrable. Let furthermore Y be a random variable with density $g(y) \propto f_1(y) f_2(y)$, then Y has truncated Gaussian distribution

$$Y \sim \mathcal{N}(\mu, \Sigma, S_1 \cap S_2)$$

where

$$\begin{pmatrix} \mu \\ \Sigma \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \Sigma_1 \end{pmatrix} * \begin{pmatrix} \mu_2 \\ \Sigma_2 \end{pmatrix}$$

and this result easily extends to any finite number of random truncated (or not) Gaussian vectors.

Lemma 9 (Conditional distribution)**.** *Let X be a random Gaussian vector in $\mathbb{R}^d$*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} R & S \\ S' & T \end{bmatrix}^{-1} \right)$$

and X_1 and X_2 the projections of X over its d_1 first and d_2 last coordinates ($d = d_1 + d_2$). The conditional distribution of X_1 with regard to X_2 is then Gaussian

$$X_1 | X_2 \sim \mathcal{N}(\mu_1 - R^{-1}S(X_2 - \mu_2), R^{-1})$$

Lemma 10. *Let X and Y be two random vectors respectively in $\mathbb{R}^d$ and $\mathbb{R}^n$ such as the conditional distribution of Y with regard to X is Gaussian*

$$Y|X \sim \mathcal{N}\left(Z + MX, \sigma^2 I_n\right)$$

with M matrix of size $n \times d$ that has full rank $d < n$, and let Z be a fixed vector in $\mathbb{R}^n$. The conditional distribution of X with regard to Y is then Gaussian too

$$X|Y \sim \mathcal{N}\left([M'M]^{-1}M'(Y - Z), \sigma^2 M'M\right).$$

Proof. Denoting $W = Y - Z$, straightforward algebra leads immediately to

$$\begin{aligned}(W - MX)'\sigma^2 I_n(W - MX) = &\left[(M'M)^{-1}M'W - X\right]'\sigma^2 M'M\left[(M'M)^{-1}M'W - X\right] \\ &- \left[(M'M)^{-1}M'W\right]'\sigma^2 M'M\left[(M'M)^{-1}M'W\right] \\ &+ W'\sigma^2 I_n W\end{aligned}$$

where the two last terms on the right hand side of the equation do not depend on X. □

6.2 MCMC algorithm for the estimation of the posterior distribution, using the informative prior

In the lines below we give the different steps of the MCMC algorithm we used to (approximately) simulate $(\theta_1, \ldots, \theta_M)$ according to the posterior distribution $\pi(\theta|y, \mathcal{D})$ corresponding to the informative prior we presented earlier. The algorithm goes as follow :

`Step 1. Initialise` θ_1 `such that` $\pi(\theta_1|y, \mathcal{D}) \neq 0$

`Step 2. For` $t = 1, \ldots, M-1$`, repeat`

`(i). Simulate` σ^2_{t+1} `cond. to` $(\alpha_t, \beta_t, \gamma_t, u_t, k_t, l_t, q_t, r_t, y, \mathcal{D})$

$$\sigma^2_{t+1} \sim \mathcal{IG}\left(\frac{N}{2}, \frac{1}{2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right)$$

`(ii). Simulate` r_{t+1} `cond. to` $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, l_t, q_t, y, \mathcal{D})$

$$r_{t+1} \sim \mathcal{G}\left(a_r + \frac{d}{2}, b_r + \frac{1}{2}\sum_{i=1}^{d}(k_i - q)^2\right)$$

`(iii). Simulate` q_{t+1} `cond. to` $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, l_t, r_{t+1}, y, \mathcal{D})$

$$q_{t+1} \sim \mathcal{N}\left([\sigma_q^{-2} + rd]^{-1}(\sigma_q^{-2} + r\sum_{i=1}^{d} k_i), [\sigma_q^{-2} + rd]^{-1}\right)$$

(iv). Simulate l_{t+1} cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, q_{t+1}, r_{t+1}, y, \mathcal{D})$

$$l_{t+1} \sim \mathcal{G}\left(a_l + \frac{d}{2}, b_l + \frac{1}{2}(\eta_t - M^{\mathcal{A}} k_t)'(\Sigma^{\mathcal{A}})^{-1}(\eta_t - M^{\mathcal{A}} k_t)\right)$$

(v). Simulate k_{t+1} cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$

$$k_{t+1} \sim \mathcal{N}\left(\mu^k_{t+1}, \Sigma^k_{t+1}\right)$$

(vi). Simulate γ_{t+1} cond. to $(\alpha_t, \beta_t, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$

$$\gamma_{t+1} \sim \mathcal{N}\left(\mu^g_{t+1}, \Sigma^g_{t+1}\right)$$

(vii). Simulate β_{t+1} cond. to $(\alpha_t, \gamma_{t+1}, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$

$$\beta_{t+1} \sim \mathcal{N}\left(\mu^b_{t+1}, \Sigma^b_{t+1}, B^{d_\beta}_+(0,1)\right)$$

(viii). Simulate α_{t+1} cond. to $(\beta_{t+1}, \gamma_{t+1}, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$

$$\alpha_{t+1} \sim \mathcal{N}\left(\mu^a_{t+1}, \Sigma^a_{t+1}\right)$$

(ix). Simulate $\delta_t \sim \mathcal{N}(0, \Sigma_{\text{MH}})$, $\quad \upsilon_t \sim \mathcal{U}[0,\, 1]$ and define $\widetilde{u}_t = u_t + \delta_t$

- define $u_{t+1} = \widetilde{u}_t$ if

$$\upsilon_t < \frac{\pi(\widetilde{u}_t | \alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})}{\pi(u_t | \alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})}$$

- or $u_{t+1} = u_t$ otherwise

where the covariance matrix Σ_{MH} used in the Metropolis-Hastings step is first estimated over a burn-in phase, and then fixed to its rescaled estimated value for the real run as in the non-informative approach.

The justifications for each full conditional distribution used in the Gibbs sampling steps, including the explicit expressions of $\mu^\alpha_{t+1}, \Sigma^\alpha_{t+1}, \mu^\beta_{t+1}, \Sigma^\beta_{t+1}, \mu^\gamma_{t+1}, \Sigma^\gamma_{t+1}, \mu^k_{t+1}$ and Σ^k_{t+1}, are now given. To derive these full conditional distributions, we will make use of the technical Lemmas 8, 9 and 10 presented earlier.

Full conditional distribution of α. The full conditional distribution of α can directly be deduced from both the prior and the likelihood contributions to it. Denote $\theta_* = (\theta, k, l, q, r)$, and write the full conditional distribution of α as

$$\pi(\alpha | \theta_* \backslash \alpha, y, \mathcal{D}) \propto g_L(\alpha) g_p(\alpha)$$

where $g_L(\alpha)$ is the contribution of the likelihood (seen as a function of α to the full conditional distribution) and $g_p(\alpha)$ is the contribution of the prior (seen as a function of α). We prove that g_L and g_p both correspond to Gaussian distributions before using Lemma 8 to combine them into yet another Gaussian distribution.

1. Let us first consider the prior contribution g_p. Recall first that α only appears in the following component of the prior

$$\pi(\theta|k,l) \propto l^{\frac{d}{2}} \exp\left(-\frac{1}{2}(\theta - M^{\mathcal{A}}k)'l(\Sigma^{\mathcal{A}})^{-1}(\theta - M^{\mathcal{A}}k)\right),$$

which directly implies that

$$g_p(\alpha) \propto \exp\left(-\frac{1}{2}(\theta - M^{\mathcal{A}}k)'l(\Sigma^{\mathcal{A}})^{-1}(\theta - M^{\mathcal{A}}k)\right).$$

Denote $\mu = M^{\mathcal{A}}k$, $\Sigma = l^{-1}\Sigma^{\mathcal{A}}$ and denote μ_α and $\mu_{\eta\backslash\alpha}$ the vectors resulting from the extractions of the coordinates corresponding to α and $\eta\backslash\alpha$ from μ. Finally denote $R_{(\alpha,\alpha)}$ the matrix resulting from the extraction of the rows and columns both corresponding to α of Σ^{-1} and denote $S_{(\alpha,\eta\backslash\alpha)}$ the one resulting from the extraction of the rows corresponding to α and columns corresponding to $\eta\backslash\alpha$ of Σ^{-1}. Using Lemma 8 (and reordering indices if necessary) it is straightforward that $g_p(\alpha)$ is proportional to the density of a Gaussian distribution

$$\mathcal{N}(\mu_\alpha - R^{-1}_{(\alpha,\alpha)}S_{(\alpha,\eta\backslash\alpha)}(\eta\backslash\alpha - \mu_{\eta\backslash\alpha}), R^{-1}_{(\alpha,\alpha)})$$

2. Let us now consider the likelihood contribution. Using exactly the same notations that we used for the full conditional distribution of α for the algorithm associated to the non-informative approach we immediately find that $g_L(\alpha)$ is proportional to the density of a Gaussian distribution

$$\mathcal{N}([M'_\alpha M_\alpha]^{-1}M'_\alpha(y - Z_\alpha), \sigma^2 M'_\alpha M_\alpha)$$

just as in (13).

3. With the help of Lemma 8 and using the two results above, we can now deduce the posterior conditional distribution of α and obtain the Gaussian distribution

$$\alpha|\theta_*\backslash\alpha, y, \mathcal{D} \sim \mathcal{N}(\mu^\alpha, \Sigma^\alpha)$$

where

$$\begin{pmatrix} \mu^\alpha \\ \Sigma^\alpha \end{pmatrix} = \begin{pmatrix} \mu_\alpha - R^{-1}_{(\alpha,\alpha)}S_{(\alpha,\eta\backslash\alpha)}(\eta\backslash\alpha - \mu_{\eta\backslash\alpha}) \\ R^{-1}_{(\alpha,\alpha)} \end{pmatrix} * \begin{pmatrix} [M'_\alpha M_\alpha]^{-1}M'_\alpha(y - Z_\alpha) \\ \sigma^2 M'_\alpha M_\alpha \end{pmatrix}.$$

Full conditional distribution of β. Using similar arguments, we obtain the full conditional distribution of β. Namely, keeping the notation introduced to derive (14), and combining the prior and the likelihood contributions together with Lemma 8 we obtain the truncated Gaussian distribution

$$\beta|\theta_*\backslash\beta, y, \mathcal{D} \sim \mathcal{N}\left(\mu^\beta, \Sigma^\beta, B^{d_\beta}_+(0,1)\right)$$

where

$$\begin{pmatrix} \mu^{\beta} \\ \Sigma^{\beta} \end{pmatrix} = \begin{pmatrix} \mu_{\beta} - R^{-1}_{(\beta,\beta)} S_{(\beta,\eta\backslash\beta)}(\eta\backslash\beta - \mu_{\eta\backslash\beta}) \\ R^{-1}_{(\beta,\beta)} \end{pmatrix} * \begin{pmatrix} [M'_{\beta}M_{\beta}]^{-1}M'_{\beta}(y - Z_{\beta}) \\ \sigma^2 M'_{\beta}M_{\beta} \end{pmatrix}.$$

Full conditional distribution of γ. Using once again similar arguments, we obtain the full conditional distribution of γ. Namely, keeping the notation introduced to derive (15), and combining the prior and the likelihood contributions together with Lemma 8 we obtain the Gaussian distribution

$$\gamma|\theta_*\backslash\gamma, y, \mathcal{D} \sim \mathcal{N}\left(\mu^{\gamma}, \Sigma^{\gamma}\right)$$

where

$$\begin{pmatrix} \mu^{\gamma} \\ \Sigma^{\gamma} \end{pmatrix} = \begin{pmatrix} \mu_{\gamma} - R^{-1}_{(\gamma,\gamma)} S_{(\gamma,\eta\backslash\gamma)}(\eta\backslash\gamma - \mu_{\eta\backslash\gamma}) \\ R^{-1}_{(\gamma,\gamma)} \end{pmatrix} * \begin{pmatrix} [M'_{\gamma}M_{\gamma}]^{-1}M'_{\gamma}(y - Z_{\gamma}) \\ \sigma^2 M'_{\gamma}M_{\gamma} \end{pmatrix}.$$

Full conditional distribution of k. Using the definition of the hierarchical prior and Lemma 8 we immediately get

$$k|\theta_*\backslash k, y, \mathcal{D} \sim \mathcal{N}\left(\mu^{k}, \Sigma^{k}\right)$$

where

$$\begin{pmatrix} \mu^{k} \\ \Sigma^{k} \end{pmatrix} = \begin{pmatrix} q\,(1, \ldots, 1)' \\ r^{-1} I_d \end{pmatrix} * \begin{pmatrix} (M^{\mathcal{A}})^{-1}\eta \\ l^{-1}\{(M^{\mathcal{A}})^{-1}\Sigma^{\mathcal{A}}(M^{\mathcal{A}})^{-1}\} \end{pmatrix}.$$

Full conditional distribution of l, q, r and σ^2). No calculations are required, as we respectively identify a gamma distribution, a Gaussian distribution, a gamma distribution, and an inverse-gamma distribution from (5).

6.3 MCMC algorithm for the estimation of the posterior distribution, using the non-informative prior

In the lines below, we give the different steps of the MCMC algorithm we used to (approximately) simulate $(\theta_1, \ldots, \theta_M)$ according to the posterior distribution $\pi(\theta|y, \mathcal{D})$ corresponding to the non-informative prior we presented earlier. The algorithm goes as follows :

`Step 1. Initialise` θ_1 `such that` $\pi(\theta_1|y, \mathcal{D}) \neq 0$

`Step 2. For` $t = 1, \ldots, M-1$`, repeat`

`(i). Simulate` σ^2_{t+1} `cond. to` $(\alpha_t, \beta_t, \gamma_t, u_t, y, \mathcal{D})$ `i.e.`

$$\sigma^2_{t+1} \sim \mathcal{IG}\left(\frac{N}{2}, \frac{1}{2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right)$$

(ii). Simulate γ_{t+1} cond. to $(\alpha_t, \beta_t, u_t, \sigma^2_{t+1}, y, \mathcal{D})$ i.e.

$$\gamma_{t+1} \sim \mathcal{N}\left(\mu^{\gamma}_{t+1}, \Sigma^{\gamma}_{t+1}\right)$$

(iii). Simulate b_{t+1} cond. to $(\alpha_t, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y, \mathcal{D})$ i.e.

$$\beta_{t+1} \sim \mathcal{N}\left(\mu^{\beta}_{t+1}, \Sigma^{\beta}_{t+1}, B^{d_\beta}_{+}(0,1)\right)$$

(iv). Simulate a_{t+1} cond. to $(\beta_{t+1}, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y, \mathcal{D})$ i.e.

$$\alpha_{t+1} \sim \mathcal{N}\left(\mu^{\alpha}_{t+1}, \Sigma^{\alpha}_{t+1}\right)$$

(v). Simulate $\delta_t \sim \mathcal{N}(0, \Sigma_{\text{MH}})$, simulate $\upsilon_t \sim \mathcal{U}[0,1]$ and define $\widetilde{u}_t = u_t + \delta_t$

- define $u_{t+1} = \widetilde{u}_t$ if

$$\upsilon_t < \frac{\pi(\widetilde{u}_t | \alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, y, \mathcal{D})}{\pi(u_t | \alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, y, \mathcal{D})}$$

- or $u_{t+1} = u_t$ otherwise

where the covariance matrix Σ_{MH} used in this last Metropolis-Hastings step is first estimated over a burn-in phase (the iterations coming from this phase are discarded), and then fixed to its estimated value "asymptotically optimally rescaled" for the final run by a factor $\left(\frac{2.38}{d}\right)^2$ (as recommended for Gaussian proposals in section 2 of Roberts and Rosenthal, 2001).

The justifications for each full conditional distribution used in the Gibbs sampling steps, including the explicit expressions of $\mu^{\alpha}_{t+1}, \Sigma^{\alpha}_{t+1}, \mu^{\beta}_{t+1}, \Sigma^{\beta}_{t+1}, \mu^{\gamma}_{t+1}$, and Σ^{γ}_{t+1}, are now given. Lemma 10 is again a key element to these justifications.

Full conditional distribution of α. The full conditional distribution of α can directly be deduced from both the prior and the likelihood contributions to it. Denote n the size of the vector α, and $\theta \backslash \alpha$ the vector θ from which the coordinates corresponding to α have been removed.

Let us first observe that, since the prior distribution we are using on α is flat, the full conditional distribution of α is in fact proportional to the likelihood function (seen as a function of α). Now considering the likelihood contribution, we write

$$\pi(\alpha | \eta \backslash \alpha, y, \mathcal{D}) \propto \exp\left(-\frac{1}{2}\sigma^{-2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right)$$

Let now L_α be the diagonal matrix whose diagonal coefficients are given by

$$(L_\alpha)_{tt} = B_{t\bullet}\beta + C_t, t = 1, \ldots, N,$$

let Z_α be the vector whose coordinates are given by

$$(Z_\alpha)_t = \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u), \quad t = 1, \dots, N,$$

and denote M_α the matrix $M_\alpha = L_\alpha A$. We can now rewrite μ and get

$$\pi(\alpha|\theta\backslash\alpha, y, \mathcal{D}) \propto \exp\left(-\frac{1}{2}\sigma^{-2}\|y - (Z_\alpha + M_\alpha\alpha)\|_2^2\right).$$

Using Lemma 10, it is then straightforward to see that the full conditional distribution of α is Gaussian

$$\alpha|\theta\backslash\alpha, y, \mathcal{D} \sim \mathcal{N}(\mu^\alpha, \Sigma^\alpha) \tag{13}$$

where

$$\begin{pmatrix} \mu^\alpha \\ \Sigma^\alpha \end{pmatrix} = \begin{pmatrix} [M'_\alpha M_\alpha]^{-1} M'_\alpha (y - Z_\alpha) \\ \sigma^2 M'_\alpha M_\alpha \end{pmatrix}.$$

Full conditional distribution of β. Using similar arguments, we obtain the full conditional distribution of β. Namely, denoting Z_β the vector whose coordinates are given by

$$(Z_\beta)_t = (A\alpha)_t C_t + \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u),$$

and calling $M_\beta = L_\beta B$ where L_β is the diagonal matrix whose diagonal is $A\alpha$, we obtain the truncated Gaussian distribution

$$\beta|\theta\backslash\beta, y, \mathcal{D} \sim \mathcal{N}(\mu^\beta, \Sigma^\beta, B_+^{d_\beta}(0, 1)) \tag{14}$$

where

$$\begin{pmatrix} \mu^\beta \\ \Sigma^\beta \end{pmatrix} = \begin{pmatrix} [M'_\beta M_\beta]^{-1} M'_\beta (y - Z_\beta) \\ \sigma^2 M'_\beta M_\beta \end{pmatrix}.$$

Full conditional distribution of γ. Using once again similar arguments, we obtain the full conditional distribution of γ. Namely, denoting Z_γ the vector whose coordinates are given by

$$(Z_\gamma)_t = (A\alpha)_t((B\beta)_t + C_t),$$

and calling M_γ the vector whose coordinates are $(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u)$ we obtain the Gaussian distribution

$$\gamma|\theta\backslash\gamma, y, \mathcal{D} \sim \mathcal{N}(\mu^\gamma, \Sigma^\gamma) \tag{15}$$

where

$$\begin{pmatrix} \mu^\gamma \\ \Sigma^\gamma \end{pmatrix} = \begin{pmatrix} [M'_\gamma M_\gamma]^{-1} M'_\gamma (y - Z_\gamma) \\ \sigma^2 M'_\gamma M_\gamma \end{pmatrix}.$$

Full conditional distribution of σ^2. No calculations are required, as we immediately identify an inverse-gamma distribution from (6).

References

Al-Zayer, J. and Al-Ibrahim, A. (1996). "Modelling the Impact of Temperature on Electricity Consumption in the Eastern Province of Saudi Arabia." *Journal of Forecasting*, 15: 97–106.

Bruhns, A., Deurveilher, G., and Roy, J. (2005). "A Non-Linear Regression Model for Mid-Term Load Forecasting and Improvements in Seasonnality." *Proceedings of the 15th Power Systems Computation Conference 2005, Liege Belgium*.

Cottet, R. and Smith, M. (2003). "Bayesian Modeling and forecasting of intraday electricity load." *Journal of the American Statistical Association*, 98(464): 839–849.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Marin, J.-M. and Robert, C. (2007). *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. Springer.

Roberts, G. and Rosenthal, J. (2001). "Optimal Scaling for various Metropolis-Hastings Algorithms." *Statistical Science*, 16(4): 351–367.

Seber, G. and Wild, C. (2003). *Nonlinear Regression*. Wiley.