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]

June 20, 2012

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 an informative hierarchical 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 working 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.

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

1 Introduction

Modelling and forecasting electricity loads is a problem well-known within both the academic and the applied statistics community (see e.g. Bunn and Farmer, 1985). The signals studied usually exhibit strong properties such as seasonalities or weekly and daily profiles, leading to some very accurate models that tend to perform rather well under normal forecasting conditions. The approaches used to model and forecast them vary a lot: we mention a couple of them in the lines below.

Some authors worked with univariate time series models: Taylor (2003) built a double seasonal exponential smoothing for the British electricity load, while Taylor et al. (2006) and Taylor and McSharry (2007) presented some comparative studies between univariate methods for different sets of data. Cugliari (2011) opted for a non-parametric approach relying on the wavelet transform to forecast the load curve seen as a functional-valued autoregressive Hilbertian process. Others tried and modelled the load together with the use of exogenous variables: Harvey and Koopman (1993) included the temperature in their model, which inspired the Bayesian semi-parametric regression model found in Smith (2000).

Alternatives to univariate modelling were often considered too, such as building multiple-equation models: while the various hours of the day share the same equation, the associated parameters differ from one another. Soares and Medeiros (2008) built an hourly independent seasonal auto-regressive model for their data, and Ramanathan et al. (1997) also built an independent model for each hour of the day but took temperature effects into account. A state-space

[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

model is proposed in Dordonnat et al. (2008) and Dordonnat (2009) where the parameters of the model are also allowed to vary along the time.

The exogenous variables most commonly used to forecast the electricity load are weather-based, even though the decision to include one or more meteorological variables into a model may be open to discussion. The period of forecasting has to be taken into account, as well as the accuracy of the predictions for such variables. For temperate climates, the most important meteorological factor is the temperature (see e.g. Taylor and Buizza, 2003). For the French electricity load specifically, the importance of the temperature and cloud-cover was underlined in Menage et al. (1988) and Bruhns et al. (2005). Other weather-related models include the works of Engle et al. (1986); Ramanathan et al. (1997) and Cottet and Smith (2003); Smith and Kohn (2002) within the Bayesian framework. Let us also mention machine learning work: Hippert et al. (2005) for a neural networks implementation and Goude (2008) for a detailed study of online mixing algorithms used on a set of various predictors.

We are interested in the development of a methodology to improve the estimation and the predictions of a parametric multi-equation model (similar to the one presented in Bruhns et al. (2005)) over a short dataset. The limited size of the dataset coupled with the high dimensionality of the model leads to a typical overfitting situation when using the maximum likelihood approach: the fitted values are relatively close to the observations while the errors in prediction are an order of magnitude larger or more (note that due to the very periodic nature of its regressors, the model typically requires 4 or 5 years of data to provide satisfactory predictions). This overfitting behaviour 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. 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.

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 an informative hierarchical (see Gelman and Hill, 2007, for a general review on the subject of hierarchical models) 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 pieces of information from this estimation (e.g. moments) to design an informative 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.

The paper is organised as follows. In Section 2 we focus on the general methodology and describe the way we carried our experimentation, 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 Section 4, we use these algorithms to illustrate and validate our approach in simulated 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 working electricity load forecasting model. In Section 5, we apply our method to French electricity datasets and compare the results with the outputs of 3 alternative standard methods to assess its competitivity. We also study the effect of the small dataset's sample size upon the predictive quality of the model and show that the informative prior provides reasonable forecasts even when the lack of data is severe.

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 and $y^{\mathcal{A}}$ the observations from $\mathcal{A}$.

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. Observe that the choice of $\pi^{\mathcal{A}}$ is not crucial as long as it remains non-informative enough because the model can be correctly estimated from the data alone on $\mathcal{A}$. Let $\pi^{\mathcal{B}}$ denote the prior distribution to be used on $\mathcal{B}$. Notice that the naive pick $\pi^{\mathcal{B}} = \pi^{\mathcal{A}}(\cdot|y^{\mathcal{A}})$ is not a viable solution as soon as the parameters of $\mathcal{A}$ and $\mathcal{B}$ differ since the variance of the posterior distribution $\pi^{\mathcal{A}}(\cdot|y^{\mathcal{A}})$ is too small: in that case the data of $\mathcal{B}$ will not be able to make up for the difference between the posterior mean on $\mathcal{A}$ and the true value of the parameters of $\mathcal{B}$. Assuming that the parameters corresponding to $\mathcal{A}$ and $\mathcal{B}$ are identical is too restrictive in practise. To allow for more flexibility we add hyperparameters accounting for the similarity between the datasets. We now described the informative hierarchical prior we designed.

Assume that the prior distribution $\pi^{\mathcal{B}}$ 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. $\lambda = T[\pi_\lambda] = F(\mathbb{E}(\theta), \dots, \mathbb{E}(\theta^m))$. The expression of $\lambda^{\mathcal{B}}$ then becomes*

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

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 informative 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 for the electricity load

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).

For each of the 48 instants of the day (each instant lasts 30 minutes, starting from 00:00AM), the non-linear regression model that we consider in this paper, first described in Bruhns et al. (2005), is made of three components, which we explain briefly in the next paragraphs, and is usually formulated as follows: for $t = 1, \dots, N$,

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

where y_t is the load of day t and where $\epsilon_1, \dots, \epsilon_N$ 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 $(\Omega_j)_{j \in \{1,\dots,d_{12}\}}$ of the calendar. This partition usually specifies

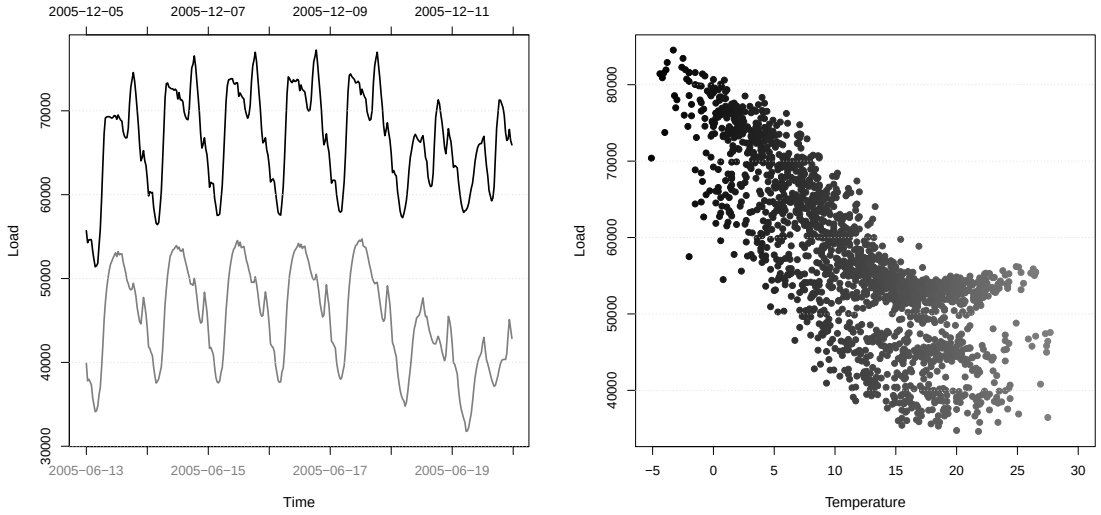

Figure 1: Left : French Electricity load from 13/06/2005 to 29/06/2005 (in grey) and from 05/12/2005 to 11/12/2005 (in black). The load is expressed in MW. Notice the daily patterns of the electricity load are not the same during summer and winter. Right : French Electricity load at 10:00 over 5 years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

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).

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 15Â°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 $\gamma \in \mathbb{R}^*$ where $\mathbb{R}^* = \mathbb{R}\backslash\{0\}$ 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.

Using the notation $M_{i\bullet}$ for the i-th row of a matrix M, the previous model can be re-written in the following condensed and more generic way: for $t = 1, \dots, N$,

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

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

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

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.

One could, without too much difficulty, add a cooling effect to the model, whose definition would be similar to that of the heating effect. Since the cooling effect remains far less important than the heating effect in France at the present time (see the right part of Figure 1), and since the estimation of the associated cooling threshold is often unstable at best, we felt that adding such a part to the model was not as crucial as it would be in other countries where the cooling effect plays a much more important role. For the applications presented in Section 5 of this paper, we thus went for a simpler implementation: given a cooling threshold u^c, a regressor whose coordinates are $(T_t - u^c)\mathbb{1}_{]-\infty, T_t]}(u^c)$, for $t = 1, \ldots, n$ is added to the matrix A. It models, in practise, a cooling effect with a fixed cooling threshold and an estimated cooling gradient that is multiplied by the daytype effect.

Considering the expression in (2), 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 rather straightforward manner, should the reader want to tweak the model a bit (for example to add another exogenous variable such as the wind or the cloud-cover).

Hereafter $L(y|\theta)$ will denote the likelihood of the observations $y = (y_1, \ldots, y_N)$. We will often write the model for $t = 1, \ldots, N$ as

$$y_t = f_t(\eta) + \epsilon_t \tag{3}$$

and use the notation $f(\eta) = (f_1(\eta), \ldots, f_N(\eta))$ for short, where $\eta = (\alpha, \beta, \gamma, u)$ designate the parameters of interest. With these notations, since $\theta = (\eta, \sigma^2)$, the likelihood of the model described in (2) reads:

$$L(y|\theta) = \left(\sqrt{2\pi}\sigma\right)^{-N} \exp\left(-\frac{1}{2}\|y - f(\eta)\|^2\right).$$

3 Specifications of the priors

3.1 Informative prior

Let us denote $\mu^{\mathcal{A}}$ and $\Sigma^{\mathcal{A}}$ the posterior mean and posterior variance of η from a non-informative approach applied to the long dataset $\mathcal{A}$, that we assume have already been collected. For the sake of clarity, we drop the $^{\mathcal{B}}$ notation: when not explicitly specified, the dataset and observations as well as the prior and posterior distributions we refer to in this Section will be those corresponding to $\mathcal{B}$.

We now present the informative hierarchical prior for the model (2), and then prove that it leads to a proper posterior distribution (see Proposition 3). Following the methodology exposed in Section 2.1, the informative prior that we propose introduces new parameters to model the similarity between the two datasets, $(k, l) \in \mathbb{R}^d \times \mathbb{R}$ and $(q, r) \in \mathbb{R} \times \mathbb{R}_+^*$ such that

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

where $K = \text{diag}(k)$. 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:

$$l \sim \mathcal{G}(a_l, b_l), \qquad q \sim \mathcal{N}(1, \sigma_q^2), \qquad r \sim \mathcal{G}(a_r, b_r), \tag{4}$$

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.

In the end, the informative prior is built as follows:

$$\pi(\theta, k, l, q, r) \propto \pi(\eta|k, l)\pi(k|q, r)\pi(l)\pi(q)\pi(r)\pi(\sigma^2) \tag{5}$$

with

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

Recalling the notations introduced at the end of Section 2, the posterior measure is given by

$$\begin{aligned}
\pi(\theta, k, l, q, r|y) &\propto L(y|\theta)\pi(\theta, k, l, q, r) \\
&\propto \sigma^{-N-2} \exp\left(-\frac{1}{2}\sigma^{-2}\|y - f(\eta)\|_2^2\right) \mathbb{1}_{[0,1]\times[\underline{u},\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1, u, \sigma^2) \\
&\quad \times l^{\frac{d}{2}} \exp\left(-\frac{1}{2}(\eta - K\mu^{\mathcal{A}})' l (\Sigma^{\mathcal{A}})^{-1}(\eta - K\mu^{\mathcal{A}})\right) \\
&\quad \times |r|^{\frac{d}{2}} \exp\left(-\frac{1}{2} r \sum_{i=1}^{d} (k_i - q)^2\right) l^{a_l - 1} \exp\left(-b_l l\right) \mathbb{1}_{\mathbb{R}_+^*}(l) \\
&\quad \times |\sigma_q^{-2}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sigma_q^{-2}(q-1)^2\right) r^{a_r - 1} \exp\left(-b_r r\right) \mathbb{1}_{\mathbb{R}_+^*}(r).
\end{aligned} \tag{6}$$

Proposition 3. *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* (6) *is then a well-defined (proper) probability distribution.*

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

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

for almost every y and that the function $\theta \mapsto \|y - f(\eta)\|_2^{-N}$ 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 prior

We now propose 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: the Bayes estimator is expected to converge, as the number of observations grows, to the maximum likelihood estimator of the model we use. The main issue when studying the asymptotical behaviour of the posterior distribution or the maximum likelihood estimator is that the likelihood of the model is not continuously differentiable with regard to the heating threshold. Isolating the heating part from the rest of the model, we show in Launay et al. (2012) that this issue can be dealt with and proove among other results that both the Bayes estimator and the maximum likelihood estimator are consistent. The non-informative prior is thus to be considered hereafter as an equivalent to the maximum likelihood approach for all intents and purposes.

For the sake of clarity again, we drop the $^{\mathcal{A}}$ notation: when not explicitly specified, the dataset and observations as well as the prior and posterior distributions we refer to in Section 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 4).

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

$$\begin{aligned} \pi(\theta|y) &\propto L(y|\theta)\pi(\theta) \\ &\propto \sigma^{-N-2} \exp\left(-\frac{1}{2}\sigma^{-2}\|y - f(\eta)\|_2^2\right) \mathbb{1}_{[0,1]\times[\underline{u},\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1, u, \sigma^2). \end{aligned} \tag{7}$$

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), *the posterior measure* (7) *is then a well-defined (proper) probability distribution.*

Proof. Notice first that

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

and observe then that

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

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

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

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

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

The triangular inequality applied to the right hand side of (8) gives

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

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 (9) as $\|\alpha_*\|_2 \to +\infty$, which is

$$\big| \|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2 \big|^{-N} \sim \lambda^{-N/2} \|\alpha_*\|_2^{-N}. \tag{10}$$

Combining (8), (9) and (10) together, we see that the integrability of the left hand side of (8) 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 $\|y - f(\eta)\|_2^{-N}$ 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 $\|y - f(\eta)\|_2^{-N}$ over $\{(\beta, u) \in B_+^{d_\beta}(0,1), \|\alpha_*\| \in]M, +\infty[\}$.

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

Remark 5. *The condition "$A_*'A_*$ has full rank" mentioned above is typically verified in our applications for the regressors used in our 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 aforementioned 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 dependent over the subset Ψ_i of the calendar".*

4 Numerical evaluations of the performance on simulated data

In this Section we simulate a long dataset $\mathcal{A}$ and a short dataset $\mathcal{B}$ from the model (2) to assess the performance of the informative prior as the similarity between the datasets varies. To measure the improvement brought by the informative prior we compare the estimation and prediction on dataset $\mathcal{B}$ with a non-informative prior. 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 informative 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 $\pi(\theta|y)$ related to the past observations $y = (y_1, \ldots, y_N)$, the predictive distribution for the future observation y_{N+k} is defined as

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

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

$$\widehat{y}_{N+k} := \mathbb{E}^{\pi}[y_{N+k}|y] \tag{11}$$

$$= \int y_{N+k} g(y_{N+k}|y)\,\mathrm{d}y_{N+k}. \tag{12}$$

The comparison criterion. To assess the quality of the estimation of the model with our informative 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 upcoming observation, the prediction error can be written as

$$y_{N+1} - \widehat{y}_{N+1} = [y_{N+1} - f_{N+1}(\eta_0)] + [f_{N+1}(\eta_0) - \widehat{y}_{N+1}],$$

which expresses the prediction error as a sum of a noise $y_{N+1} - f_{N+1}(\eta_0)$ (whose theoretical distribution is $\mathcal{N}(0, \sigma^2)$) and a bias which can be seen as an estimation error over the prediction $f_{N+1}(\eta_0) - \widehat{y}_{N+1}$. 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 $f_{N+1}(\eta_0) - \widehat{y}_{N+1}$ 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.:

$$\sqrt{\frac{1}{365}\sum_{i=1}^{365}[f_{N+i}(\eta_0) - \widehat{y}_{N+i}]^2}. \tag{13}$$

4.2 Construction of simulated datasets

Both datasets $\mathcal{A}$ and $\mathcal{B}$ were simulated according to the model (2) 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 bank holidays. 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:

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

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 informative 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 informative prior against its 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 informative prior (most of them hitting the true value). The marginal posterior standard deviations (not shown here) are also reduced when the informative prior is used instead of the non-informative prior.

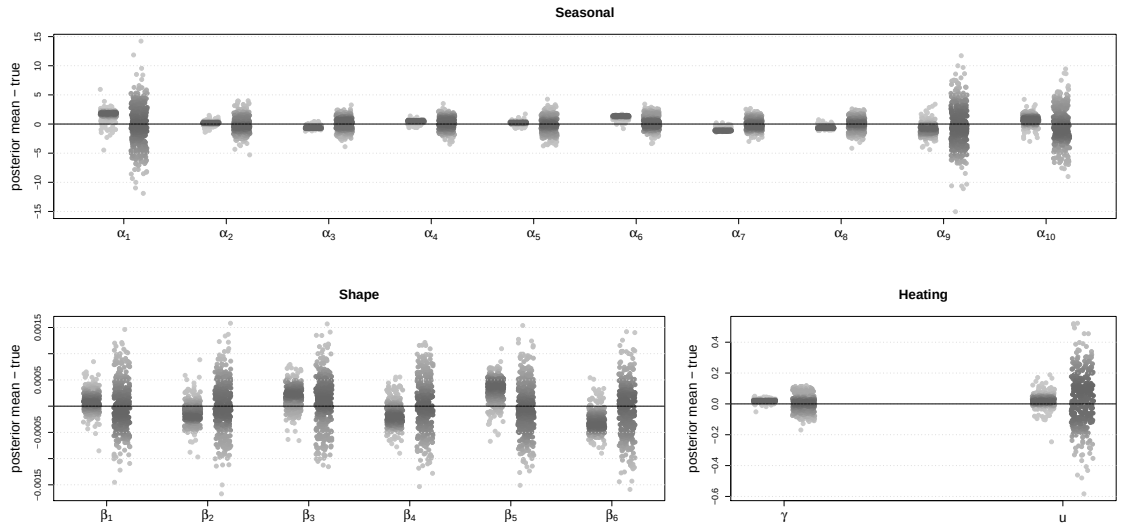

Figure 2: The posterior error (posterior mean minus true value) of α (seasonal parameters), β (shape parameters), and γ and u (heating parameters), based on 300 replications. Leftmost replications correspond to the informative method while the rightmost replications correspond to the non-informative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

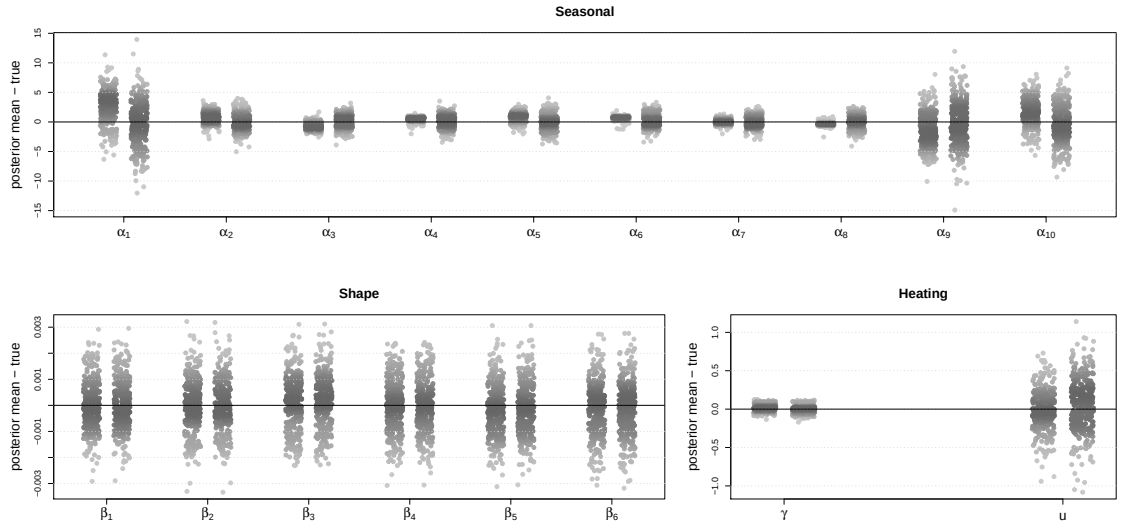

Figure 3: Same caption as in Figure 2 except $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$.

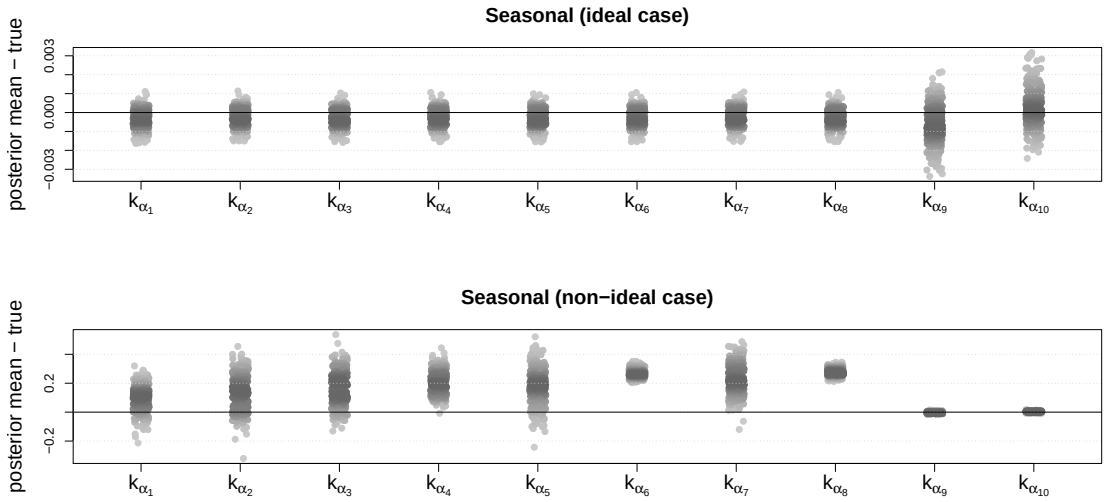

Figure 4: The posterior error (posterior mean minus true value) of k_α (seasonal parameters), based on 300 replications. Top row is for the case were $k_\alpha = k_\beta = k_\gamma = k_u = 1$ and bottom row is for the case where $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$. Leftmost replications correspond to the informative 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).

When the situation is far from being as ideal as the one mentioned above, the informative 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, 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 informative 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

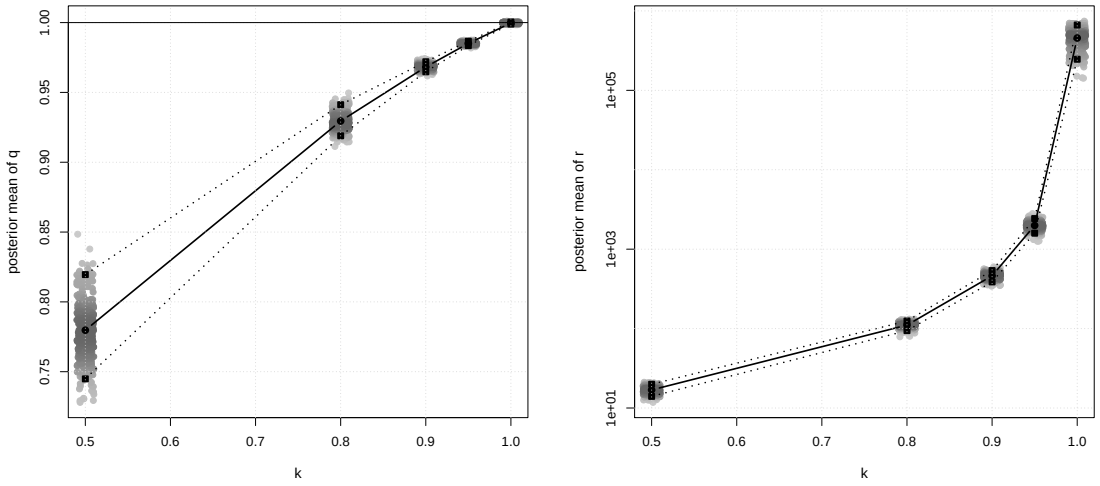

Figure 5: In grey: posterior mean of q (left) and r (right, on a log scale) for the informative prior (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $k_\alpha = k_\gamma$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

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 informative prior in (5) 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 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 (5) 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 informative and the non-informative models using our comparison criterion defined in (13) and computing the ratio between the two models for different values of k_α and k_γ, k_β and k_u being both set 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 included one side of the graph in the present article.

On average, the Bayesian informative 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

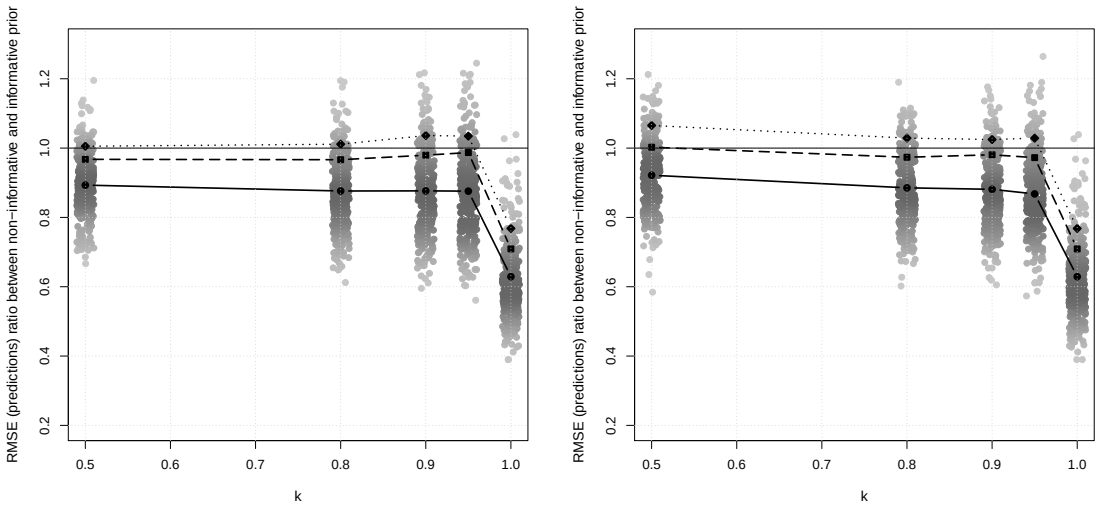

Figure 6: In grey: ratio between error predictions for the informative and the non-informative approach (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $k_\alpha = k_\gamma$ (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.

of the informative hierarchical prior 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 Applications

The long dataset $\mathcal{A}$ needed for the construction of the informative prior corresponds to a specific population in France frequently referred to as "non-metered" because their electricity consumption is not directly observed by EDF but instead derived as the difference between the overall electricity consumption and the consumption of the "metered" population. For this population the data ranged from 07/01/2004 to 07/31/2010. We illustrate the benefit of choosing our informative prior to predict electricity load on short datasets. We consider two short datasets : the first $\mathcal{B}$ corresponds to the "non-metered" population for ERDF, a wholly owned subsidiary of EDF that manages the public electricity network for 95% of continental France. This population roughly covers the same people that $\mathcal{A}$ does, but not exactly. The second dataset $\mathcal{B}'$ corresponds to a subpopulation of $\mathcal{A}$ and represents around 50% of the total load of $\mathcal{A}$.

5.1 Benchmark against standard methods

For this application, only the days for which no special tariffs are enforced were considered: the so-called EJP ("Effacement jour de pointe" = peak tariff days) were removed from the dataset beforehand to ensure the signal studied was consistent throughout time. Bank holidays (including the day before and the day after to avoid any neighbourhood contamination effects), the summer holiday break (August) and the winter holiday break (late December) were also removed from

	Estimation $\mathcal{A}$	Estimation $\mathcal{B}$	Prediction $\mathcal{B}$
Case 1	1099	125	28
Case 2	1099	144	38

Table 1: Sample size (in days) of the datasets for both experiments.

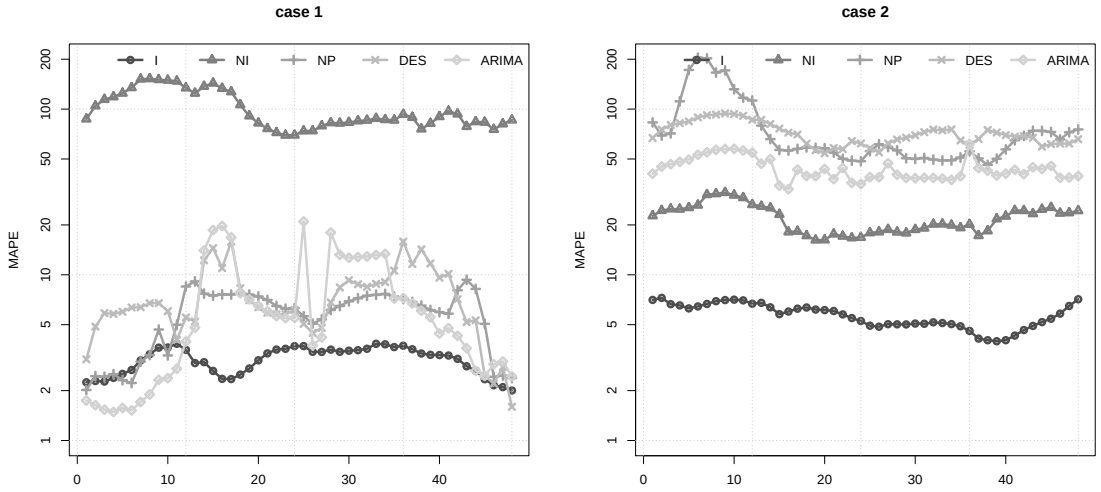

Figure 7: Quality of the predictions broken by instant (MAPE in %), for both experiments: case 1 (left) and case 2 (right). The ordinate axis is in log-scale. Each shade of gray corresponds to one of the 5 methods tested.

the dataset for this first application, as we wanted to benchmark our method against others on a smooth and rather easy-going signal. The temperature considered in the model is the average temperature over France for the period of study, and the cooling threshold was chosen to be 16Â°C throughout the 48 instants of the day.

We benchmarked our Bayesian method with informative prior against four alternative methods, comparing their predictions on dataset $\mathcal{B}$ in two configurations. Roughly speaking, for our first experiment we estimated the model for $\mathcal{B}$ over the period ranging from 12/01/2009 to 06/30/2010 and predicted the next 30 days (same as the application presented Section 5.2), while for our second experiment, we estimated the model for $\mathcal{B}$ over the period ranging from 01/01/10 to 07/31/10 and predicted the previous 30 days. We expect the first configuration to be the easy case and the second configuration to be the tough case, the signal being very smooth during summer and not so much during winter. The figures shown in Table 1 summarise the exact lengths of the various datasets for both experiments.

The four alternative methods we benchmarked against our own Bayesian informative method, relied on four different techniques: the first was the Bayesian non-informative method that we exposed earlier in Section 3.2 (recall that it was meant to be an equivalent to the maximum likelihood approach), the second involved non-parametric estimation with kernels (see Fan and Yao, 2005), the third was a double exponential smoothing (see Taylor, 2003) and the fourth and last was an ARIMA model. Note that for the second experiment, the data available obviously had to be time-reversed in order to apply some the last three alternatives methods since time-dependence plays an important role for them. The ARIMA model was automatically selected (see Hyndman and Khandakar, 2008) and was the best model with regard to the AIC criterion.

	Case 1		Case 2	
	RMSE	MAPE	RMSE	MAPE
informative prior (I)	770.71	3.08	2041.70	5.71
non-informative prior (NI)	24440.52	100.26	7317.56	22.03
non-parametric (NP)	1461.52	5.82	25091.68	77.41
double exp. smoothing (DES)	1800.12	7.41	22572.58	71.36
linear time series (ARIMA)	1702.39	6.89	13839.85	43.73

Table 2: Overall quality (RMSE in MW, MAPE in %) of the predictions for both experiments.

It is clear from the results exposed in Table 2 that the informative prior outperforms all the alternative methods by a large margin in each case. Figure 7 shows that the Bayesian informative method is superior to the Bayesian non-informative method, the non-parametric approach, the double exponential smoothing method as well as the ARIMA Model throughout the 48 instants of the day in both cases, with the exception of night-time for Case 1 where the non-parametric and ARIMA model remain competitive. This can be attributed to the nocturnal signal being very smooth in July, compared to the signal in winter. The overall bad performance of the Bayesian non-informative method is not surprising because at least 3 to 4 years of data are usually required to avoid overfitting, for such a parametric model.

5.2 Role of the hyperparameters

For this application, the setup was the same as the one described at the start of Section 5.1. Our aim was to point out the role of the hyperparameters introduced within the informative prior and show that besides providing better results than alternative methods as was demonstrated in Section 5.1, they also provided a measure of similarity between the short datasets of interest and the dataset used to build the prior. We estimated the model for $\mathcal{B}$ and $\mathcal{B}'$ over the period ranging from 12/01/2009 to 06/30/2010 and predicted the next 30 days.

Figure 8 shows the predictive quality of the model for both populations $\mathcal{B}$ and $\mathcal{B}'$ using the informative prior as well as the posterior densities of the similarity coefficients k_j at midday (we also looked at the other 47 instants but the look of them was nearly identical to the one we chose to present). These densities are much more peaked and also centred closer to 1 for population $\mathcal{B}$ than they are for population $\mathcal{B}'$. It thus seems to indicate that dataset $\mathcal{B}$ is more similar to $\mathcal{A}$ than $\mathcal{B}'$ is, confirming our prior knowledge that $\mathcal{A}$ and $\mathcal{B}$ covered approximately the same population whereas $\mathcal{B}'$ represented around 50% of $\mathcal{A}$: this value of 50% is also visible on Figure 8 where we observe two densities centred around 0.5, which correspond to the similarity coefficients between the offsets ω_j of $\mathcal{A}$ and these of $\mathcal{B}'$.

Figure 9 displays the boxplots for the posterior densities of q and $1/\sqrt{r}$ and seem to corroborate the fact that $\mathcal{B}$ is more similar to $\mathcal{A}$ than $\mathcal{B}'$ is. Recall that q and $1/\sqrt{r}$ respectively act as the mean and standard deviation of the similarity coefficients k_j within our informative hierarchical prior. Indeed the estimated mean of q appears to be closer to 1 while its estimated variance is smaller on $\mathcal{B}$ than $\mathcal{B}'$. The estimated mean and variance of $1/\sqrt{r}$ are also smaller on $\mathcal{B}$ than $\mathcal{B}'$.

As we observed in Section 4 when we dealt with simulated datasets, the estimated values of q and r provide some information about the similarity between the datasets considered. Notice also that, here again, the best predictive performance is obtained when the similarity between the two datasets is strongest.

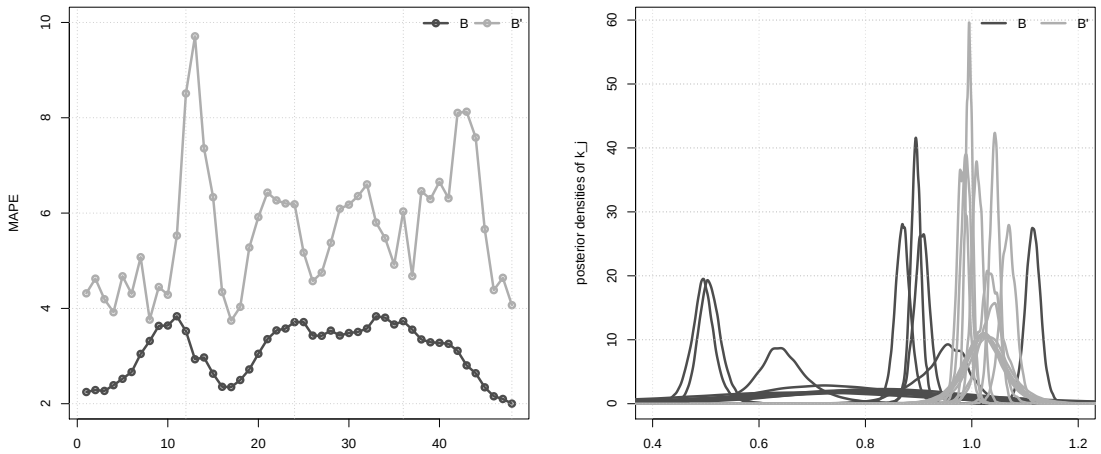

Figure 8: Quality of the predictions broken by instant (MAPE in %), for populations $\mathcal{B}$ and $\mathcal{B}'$ (left). Posterior densities of the similarity coefficients k_j for populations $\mathcal{B}$ and $\mathcal{B}'$ at midday (right).

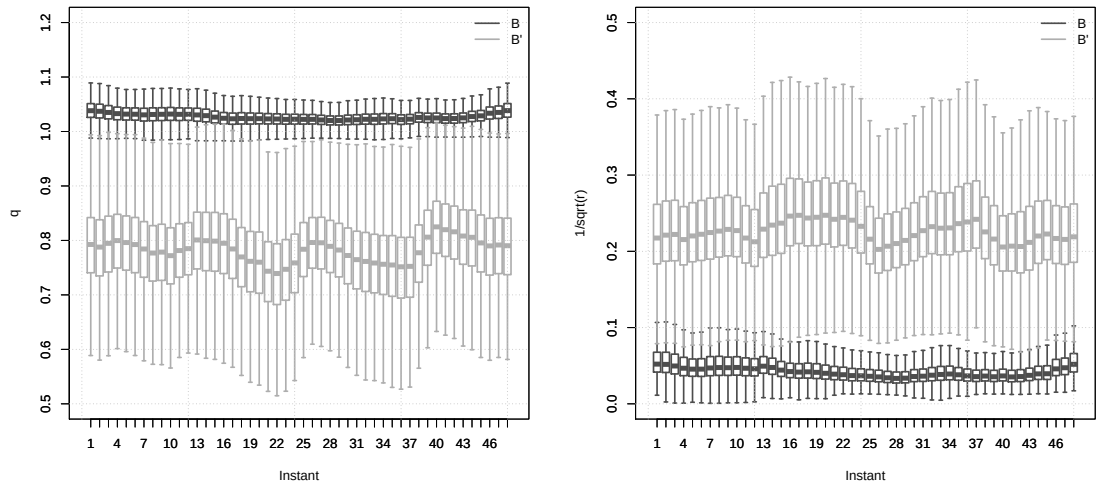

Figure 9: Boxplots of the posterior densities of q (left) and $1/\sqrt{r}$ (right) (mean and standard deviation of the similarity coefficients k_j) for populations $\mathcal{B}$ and $\mathcal{B}'$, throughout the 48 instants of the day.

5.3 Role of the sample size

For this application, the setup is almost identical to the one described at the start of Section 5.1 but the temperature considered in the model is not the average temperature over France anymore but a transformation of it: it was smoothed using exponential smoothing, which is known to improve the link existing between the two variables temperature and electricity load (see Bruhns et al., 2005, for more information about this). The cooling threshold was fixed at 18Â°C throughout the 48 instants of the day, and this time, the summer holiday break was not removed from the dataset (but the winter holiday break and the bank holidays still were), so that the model could benefit from (and be tested on) the August months in general.

We put the focus on the length of the estimation period on $\mathcal{B}$ while keeping the same prediction window. We successively chose the periods ranging from 01/01/2010, 03/01/2010, 05/01/2010, 07/01/2010 to 12/31/2010, reducing the estimation period on $\mathcal{B}$ from 12 months to only 6 months, removing 2 months at a time. The next 6 upcoming months were then predicted i.e. the prediction window ranged from 01/01/2011 to 06/30/2011. Note that for $\mathcal{A}$ we used the same dataset that we used for our two first applications. The diagram in Figure 10 describes the 4 scenarios considered.

The non-informative prior leads to a better fit than the informative prior as can be seen in Table 3. It should not come as a surprise because the non-informative prior was indeed meant to be equivalent to a maximum likelihood approach whose criterion is precisely to minimise the RMSE. As for the quality of the predictions associated with the model for both priors, Table 3 demonstrates that the informative prior beats the non-informative prior in each of the four proposed configurations. The improvement appears to be minimal when 12 months are used but as months are removed from the estimation window, the predictive quality for the non-informative prior drops very quickly, while the predictive quality for the informative prior remains moderate and stable.

Figure 11 shows the quality of the predictions broken by months while Figures 12 and 13 display the same, but broken by instants. It is important to note that the use of the non-informative prior leads to results exhibiting all the typical signs of overfitting a model. The bias induced by the increasing lack of data is mainly seasonal, as can be seen on Figure 11: this is due to the seasonality coefficients of the model being overfitted. Choosing the informative prior over the non-informative prior makes the estimation and prediction of the model more robust with regards to the lack of data.

The informative prior especially improves the quality of the predictions when the lack of data is severe: it provides reasonable forecasts even in the worst scenario considered here, where only 6 months of estimation were used for 6 months of prediction. In this situation, estimation (from 07/01/2010 to 12/31/2010) and prediction (from 01/01/2011 to 06/30/2011) are performed on non-overlapping areas of the calendar: the informative prior makes up for the unavailable data and prevents the model from overfitting on the second half of the calendar, while the non-informative prior does not and consequently leads to heavily biased predictions over the first half of the calendar.

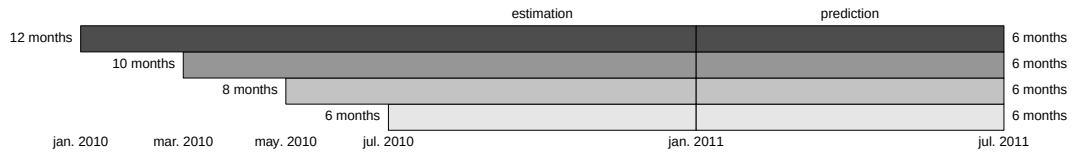

Figure 10: Ranges of the estimation (from 12 to 6 months) and prediction (6 months) time-windows for the 4 scenarios considered.

	Estimation				Prediction			
	RMSE		MAPE		RMSE		MAPE	
	non-info.	info.	non-info.	info.	non-info.	info.	non-info.	info.
12m.	663.02	671.95	1.86	1.87	763.23	737.83	2.01	1.94
10m.	606.04	623.23	1.78	1.82	1509.09	883.07	3.18	2.21
8m.	473.29	493.68	1.49	1.52	8891.81	1318.28	16.72	3.26
6m.	460.60	499.13	1.34	1.44	90356.82	1305.27	224.40	3.62

Table 3: Overall quality (RMSE in MW, and MAPE in %) of the estimation (left) and prediction (right) for the non-informative (non-info.) and informative (info.) priors, depending on the number of months used for the estimation (from 12 months to 6 months).

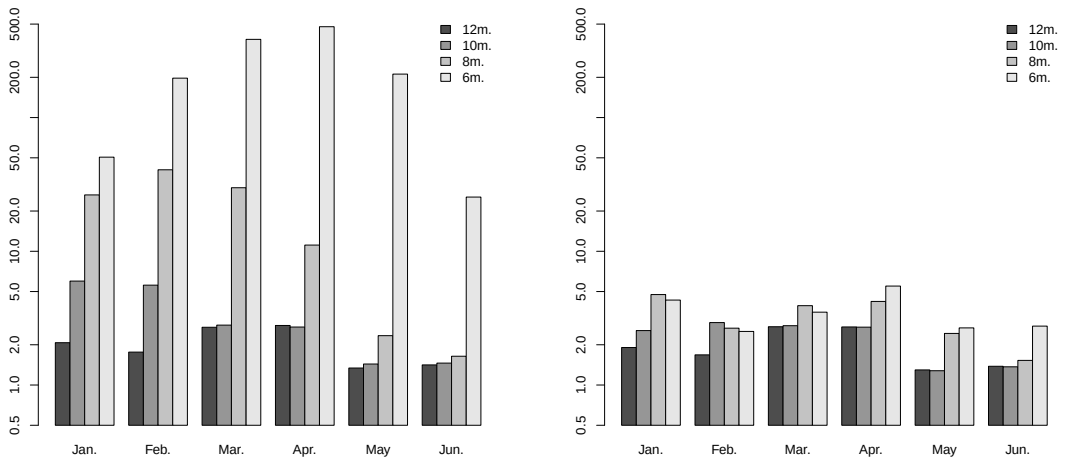

Figure 11: Quality of the predictions broken by month (MAPE in %), with an estimation period ranging from 12 to 6 months using the non-informative prior (left) and using the informative prior (right). The ordinate axis is in log-scale. Each shade of gray corresponds to a different scenario.

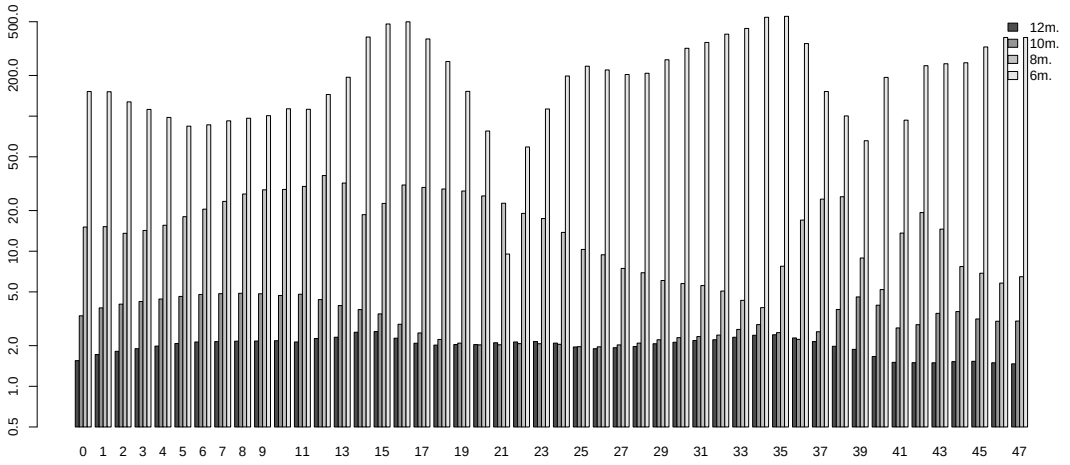

Figure 12: Quality of the predictions broken by instant (MAPE in %), with an estimation period ranging from 12 to 6 months using the non-informative prior. The ordinate axis is in log-scale. Each shade of gray corresponds to a different scenario.

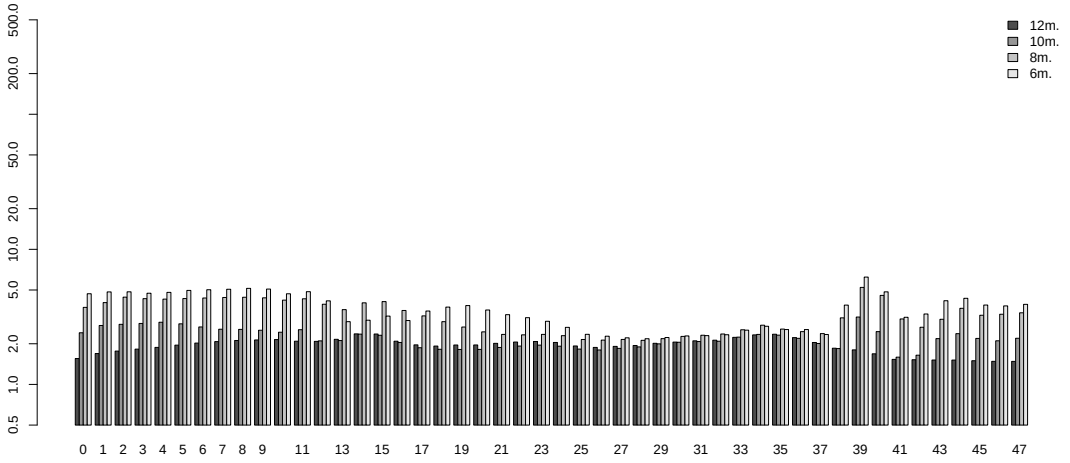

Figure 13: Quality of the predictions broken by instant (MAPE in %), with an estimation period ranging from 12 to 6 months using the informative prior. The ordinate axis is in log-scale. Each shade of gray corresponds to a different scenario.

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 used a Metropolis-Hasting step based on a Gaussian random walk proposal. The algorithm corresponding to the non-informative prior is detailed first since it is the simplest of the two.

6.1 Technical Lemmas

Definition 6 (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 7 (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 8 (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 9. *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 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 $\pi(\theta|y)$ corresponding to the non-informative prior we presented earlier. The algorithm goes as follows:

Step 1. Initialise θ_1 such that $\pi(\theta_1|y) \neq 0$

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

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

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

(ii). Simulate γ_{t+1} cond. to $(\alpha_t, \beta_t, u_t, \sigma^2_{t+1}, y)$ 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 $(\alpha_t, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y)$ 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 $(\beta_{t+1}, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y)$ 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 $v_t \sim \mathcal{U}[0,1]$ and define $\widetilde{u}_t = u_t + \delta_t$

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

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

- 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 9 is a key element to these justifications.

Full conditional distribution of α. 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. The full conditional distribution of α can directly be deduced from both the prior and the likelihood contributions to it.

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

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

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

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

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

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

Using Lemma 9, it is then straightforward to see that the full conditional distribution of α is Gaussian

$$\alpha|\theta\backslash\alpha, y \sim \mathcal{N}(\mu^\alpha, \Sigma^\alpha) \tag{14}$$

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

$$\beta|\theta\backslash\beta, y \sim \mathcal{N}(\mu^\beta, \Sigma^\beta, B_+^{d_\beta}(0,1)) \tag{15}$$

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 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

$$\gamma|\theta\backslash\gamma, y \sim \mathcal{N}(\mu^\gamma, \Sigma^\gamma) \tag{16}$$

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 (7).

6.3 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 $\pi(\theta|y)$ corresponding to the informative prior we presented earlier. The algorithm goes as follow:

`Step 1. Initialise` θ_1 `such that` $\pi(\theta_1|y) \neq 0$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- 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 7, 8 and 9 presented earlier.

Full conditional distribution of α. 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. 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

$$\pi(\alpha | \theta_* \backslash \alpha, y) \propto g_L(\alpha) g_p(\alpha)$$

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 7 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

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

which directly implies that

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

Denote $\mu = K\mu^{\mathcal{A}}$, $\Sigma = l^{-1}\Sigma^{\mathcal{A}}$ 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 7 (and reordering indexes 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 with 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 (14).

3. With the help of Lemma 7 and using the two results above, we can now deduce the posterior conditional distribution of α and obtain the Gaussian distribution

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

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 (15), and combining the prior and the likelihood contributions together with Lemma 7 we obtain the truncated Gaussian distribution

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

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 again similar arguments, we obtain the full conditional distribution of γ. Namely, keeping the notation introduced to derive (16), and combining the prior and the likelihood contributions together with Lemma 7 we obtain the Gaussian distribution

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

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. We denote $M^{\mathcal{A}} = \text{diag}(\mu^{\mathcal{A}})$ and first notice that $M^{\mathcal{A}} k = K\mu^{\mathcal{A}}$. Using the definition of the informative prior and Lemma 7, we then immediately derive

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

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 (6).

Acknowledgments

The authors would like to thank Adélaïde Priou for collecting a part of the data as well as the corresponding results, and Virginie Dordonnat for the insightful discussions.

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*.

Bunn, D. and Farmer, E. (1985). Comparative models for electrical load forecasting.

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

Cugliari, J. (2011). *Prevision non parametrique de processus a valeurs fonctionnelles. Application a la consommation d'electricite*. PhD thesis, Universite Paris Sud XI.

Dordonnat, V. (2009). *State-space modelling for high frequency data*. PhD thesis, Vrije Universiteit Amsterdam.

Dordonnat, V., Koopman, S., 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.

Engle, R., Granger, C., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity. *Journal of the American Statistical Association*, 81:310–320.

Fan, J. and Yao, Q. (2005). *Non linear Time Series: Nonparametric and Parametric Methods*. Springer.

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

Goude, Y. (2008). *Melange de predicteurs, application a la prevision de consommation d'electricite*. PhD thesis, Universite Paris Sud XI.

Harvey, A. C. and Koopman, S. J. (1993). Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association*, 88:1228–1237.

Hippert, H., Bunn, D., and Souza, R. (2005). Large neural networks for electricity load forecasting: Are they overfitted? *International Journal of Forecasting*, 21:425–434.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 88(3).

Launay, T., Philippe, A., and Lamarche, S. (2012). Consistency of the posterior distribution and MLE for piecewise linear regression. *Preprint*. arXiv:1203.4753.

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

Menage, J. P., Panciatici, P., and Boury, F. (1988). Nouvelle modelisation de l'influence des conditions climatiques sur la consommation d'energie electrique. Technical report, EDF R&D.

Ramanathan, R., Engle, R., Granger, C., Vahid-Araghi, F., and Brace, C. (1997). Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting*, 13:161–174.

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.

Smith, M. (2000). Modeling and short-term forecasting of new south wales electricity system load. *Journal of Business & Economic Statistics*, 18:465–478.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97:1141–1153.

Soares, L. and Medeiros, M. (2008). Modeling and forecasting short-term electricity load: a comparison of methods with an application to brazilian data. *International Journal of Forecasting*, 24:630–644.

Taylor, J. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54:799–805.

Taylor, J. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19:57–70.

Taylor, J., De Menezes, L., and McSharry, P. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, 22:1–16.

Taylor, J. and McSharry, P. (2007). Short-term load forecasting methods: An evaluation based on european data. *IEEE Transactions on Power Systems*, 22:2213–2219.