HAL
archives-ouvertes.fr

Construction of an informative hierarchical prior for a small sample with the help of historical data and application to electricity load forecasting

Tristan Launay, Anne Philippe, Sophie Lamarche

▶ To cite this version:

HAL Id: hal-00625117

https://hal.archives-ouvertes.fr/hal-00625117v5

Submitted on 25 Mar 2014

Construction of an informative hierarchical prior for a small sample with the help of historical data and application to electricity load forecasting

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

March 25, 2014

Abstract

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 illustrate the performance of our prior on simulated dataset from three standard models. Then we apply the methodology to a working model for the electricity load forecasting on 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

We consider the problem of building a prior distribution when only a small number of observations is available. In this context the choice of the prior becomes crucial, having a strong effect on the inference. However it can also be a good tool to bring information so as to improve the estimations in complicated problems (e.g. model in high dimension with respect to the number of observations). Using the ideas of hierarchical modelling, we propose a parametric prior model, that requires the knowledge of another long historical sample $y^{\mathcal{H}}$ upon which the model performs equally well in both estimation and prediction. We will assume that this historical sample shares some similarities with our short sample of interest y.

The study is motivated by the modelling and forecasting of electricity load which is a problem well-known within both the academic and the applied statistics community (see e.g. Bunn and Farmer, 1985). 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 (due to the very periodic nature of the regressors involved, 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

---

[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

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.

Different strategies can be found in the literature of hierarchical model to choose or to estimate the parameters of the prior distribution. See Gelman et al. (2013); Gelman and Hill (2007); Congdon (2010) for a general review on the subject of hierarchical models. These parameters can be assumed unknown and a new level is added to the hierarchical structure. This commonly used approach increases the dimension of the model, which is generally not recommended when dealing with small samples.

In Gelman et al. (2013), Chapter 5, some methods using historical samples are presented and illustrated on several examples. Typically, both samples are supposed to be coming from a common model. Assuming that the prior belongs in a parametric family, the hyperparameters of the prior are chosen as the estimated values on the historical sample. This strategy is not robust to small evolutions of the parameters between the historical and the short samples, leading to a bias in the estimation as illustrated in Section 4. Note that it is also possible to estimate the prior on the sample itself, the approaches are often called empirical Bayesian method. This is not a full Bayesian analysis but the approach can be viewed as an approximation to hierarchical models (see e.g. Efron (2010) for details).

Instead of estimating the parameter from historical sample, one can define a joint model on the samples $(y^{\mathcal{H}}, y)$ where both samples come from the same parametric model but have different parameters (denoted respectively $\theta_{\mathcal{H}}$ and θ). The prior distribution on $(\theta_{\mathcal{H}}, \theta)$ conveys the link and the closeness between both parameters with the help of a dependence structure. Bouveyron and Jacques (2013) propose this method to estimate regression mixtures when the population has changed between the training and the predictive stages. They choose a conditional distribution of θ given $\theta_{\mathcal{H}}$ which is centered in $\theta_{\mathcal{H}}$. This approach requires knowledge of the joint distribution of samples $(y^{\mathcal{H}}, y)$, which are assumed independent in Bouveyron and Jacques (2013). This leads to a transfer learning in both directions and attributes a higher importance to the largest historical sample.

In this paper, we assume that 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 we consider) 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 prior for the short dataset from the long historical 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 are enough to estimate and predict the model in this case ;

2. we extract key pieces of information from this estimation 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 propose a informative Gaussian prior model. From the two first moments, we show the transfer of information between the historical sample and the prior model. In Section 3 we illustrate the performance of our methodology on simulated dataset from three standard models : the auto-regression model, the polynomial regression and a Poisson model with random effects. In Section 4, we introduce the model used for our main application of electricity load forecasting. We

discuss the results obtained on French electricity datasets and show that our new methodology is competitive compared to three alternative standard methods. 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. The ad hoc MCMC algorithms we developed to approximate numerically the various estimators are push backed into the Appendix (see Section 5) so as not to obfuscate the main point of the paper by technical details.

2 Methodology

2.1 General principle

We denote y the short sample over which we would like to estimate a parametric model $\{f_\theta, \theta \in \Theta\}$ and to predict future observations. We denote $y^{\mathcal{H}}$ a historical sample coming from the same parametric model. We assume that this is a long dataset known to share some common features with y. Let $\theta_{\mathcal{H}}$ and θ be the corresponding parameters. The parameter of interest is θ .

Let $\pi_{\mathcal{H}}$ and π be the prior distributions used to estimate $\theta_{\mathcal{H}}$ and θ respectively. Let $\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})$ and $\pi(\cdot|y)$ be the associated posterior distributions. We will denote $\widehat{\theta}_{\mathcal{H}}$ and $\widehat{\theta}$ the corresponding Bayes estimates (i.e. the posterior means).

We propose a method to the improve parameter the estimations and model predictions for the dataset y with the help of the dataset $y^{\mathcal{H}}$. Observe first that the choice of $\pi_{\mathcal{H}}$ is not crucial as long as it remains non-informative enough because, $\theta_{\mathcal{H}}$ can be correctly estimated from the data alone on $y^{\mathcal{H}}$. The naive pick $\pi = \pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})$ is not a viable solution as soon as the parameters of $\theta_{\mathcal{H}}$ and θ differ since the variance of the posterior distribution $\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})$ is too small: in that case the few observations of y will not be able to make up for the difference between the posterior mean $\widehat{\theta}_{\mathcal{H}}$ and the true value of the parameter θ. Assuming that the parameters $\theta_{\mathcal{H}}$ and θ are identical is too restrictive in practise. To allow for more flexibility we introduce hyperparameters to account for the similarity between the datasets. We now describe the informative hierarchical prior we designed.

Assume that the prior distribution π is to be chosen within the parametric family

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

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

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

and choose λ proportional to $T[\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})]$, in the sense that

$$\lambda = KT[\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})],$$

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 $\theta_{\mathcal{H}}$ and θ, and its diagonal components as similarity coefficients measuring how close the two datasets $y^{\mathcal{H}}$ and y really are when looked at through the model. 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.

Remark 1. *As we suppose that $y^{\mathcal{H}}$ is a large sample, the posterior distribution $\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})$ can be approximated by Gaussian distribution under suitable regularity conditions. Therefore we can replace*

$$\lambda = KT[\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})],$$

by

$$\lambda = KT[\mathcal{N}(\hat{\theta}_{\mathcal{H},n}^{ML}, I_{\mathcal{H},n}^{-1})],$$

where $\hat{\theta}_{\mathcal{H},n}^{ML}$ is the maximum likelihood estimate and where $I_{\mathcal{H},n}$ is an estimate of the expected Fisher Information matrix calculated on the sample $y^{\mathcal{H}}$ (Ghosh et al., 2006, chap 4).

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

Example 2 (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 λ then becomes*

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

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 $y^{\mathcal{H}}$ remains untouched and is directly plugged into in the informative prior if we consider centred moments for orders greater than 1.

Example 3 (Conjugacy)**.** *We consider the case where $\mathcal{F}$ is the family of priors conjugated for the model. If the prior $\pi_{\mathcal{H}}$ belongs to $\mathcal{F}$ then the associated posterior distribution $\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})$ does too and we denote $\lambda^{\mathcal{H}}(y^{\mathcal{H}})$ the parameter that corresponds to it. The expression of λ then reduces to*

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

2.2 Gaussian strategy

Under suitable conditions of regularity (see Ghosh et al. (2006), Launay et al. (2012)), the posterior distribution tends to look like a normal distribution asymptotically. This suggests the use of a normal distribution as the prior distribution π. Let us denote $\hat{\mu}^{\mathcal{H}}$ and $\hat{\Sigma}^{\mathcal{H}}$ the posterior mean and posterior variance of the parameter $\theta_{\mathcal{H}}$.

As described in the general methodology, we introduce new parameters to model the similarity between the two datasets, $(k,l) \in \mathbb{R}^d \times \mathbb{R}$ and $(q,r) \in \mathbb{R} \times \mathbb{R}_+^*$. The informative hierarchical prior is of the form

$$\theta|k,l \sim \mathcal{N}(K\hat{\mu}^{\mathcal{H}}, l^{-1}\hat{\Sigma}^{\mathcal{H}}) \tag{1}$$

$$k|q,r \sim \mathcal{N}(q(1,\ldots,1)', r^{-1}I_d) \tag{2}$$

where $K = \text{diag}(k)$. Again, the coordinates of the vector k can be interpreted as similarity coefficients between $\theta_{\mathcal{H}}$ and θ 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 $\theta_{\mathcal{H}}$ and θ are, q corresponding to the mean of the coordinates of k and

r being their inverse-variance. For a complete definition of the prior distribution, we of course need to specify the prior on l, q, r. Based on conjugacy considerations, we use:

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

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 (that we took advantage of to design the MCMC algorithms for our applications). 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:

$$\pi(\theta, k, l, q, r) \propto \pi(\theta|k, l)\pi(k|q, r)\pi(l)\pi(q)\pi(r) \qquad (4)$$

where the densities appearing on the right hand side are specified in (1), (2) and (3).

From the two first moments of the prior distribution π, we evaluate the prior information brought by the historical sample $y^{\mathcal{H}}$. Firstly we show that the prior distribution π is centered on the estimation of the parameter on $y^{\mathcal{H}}$, which should be close to $\theta_{\mathcal{H}}$ according to the asymptotic properties of Bayes estimates.

Property 4. *The prior distribution admits a moment of order 1 equal to $\widehat{\mu}^{\mathcal{H}}$.*

Proof. Indeed, we have

$$\begin{aligned}
\mathbb{E}(\theta) &= \mathbb{E}(\mathbb{E}(\theta|K, l^{-1})) = \mathbb{E}(K\widehat{\mu}^{\mathcal{H}}) \\
&= \mathbb{E}(K)\widehat{\mu}^{\mathcal{H}} = \mathbb{E}(\mathbb{E}(K|q, r))\widehat{\mu}^{\mathcal{H}} = \mathbb{E}(q)I_d\widehat{\mu}^{\mathcal{H}} \\
&= \widehat{\mu}^{\mathcal{H}}.
\end{aligned}$$

$\square$

Secondly, from the correlation matrix of π, we show that we transfer informations from the historical sample to the dependence between the components of parameter θ

Property 5. *Assume that $a_l > 1$ and $a_r > 1$. Under suitable conditions of regularity,*

- *the correlation of the prior distribution is equivalent to the correlation of the posterior distribution $\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})$ as $N_{\mathcal{H}}$ goes to infinity.*

- *for a fixed $N_{\mathcal{H}}$ the correlation of the prior distribution is also equivalent to the correlation of the posterior distribution $\pi_{\mathcal{H}}(\cdot|y^{\mathcal{H}})$ as σ_q^2 and $\mathbb{E}(r^{-1})$ go to 0.*

Proof. First notice that the condition on the coefficients $a_l > 1$ and $a_r > 1$ ensures that the first moment of both l^{-1} and r^{-1} are finite. This implies that the variance of the prior distribution is finite

Recall now that if (X, Y) are two random variables in L^2 we have the following equality

$$\text{var}(X) = \text{var}(\mathbb{E}(X|Y)) + \mathbb{E}(\text{var}(X|Y)).$$

We then get

$$\begin{aligned}
\text{var}(\theta) &= \text{var}(\mathbb{E}(\theta|K, l^{-1})) + \mathbb{E}(\text{var}(\theta|K, l^{-1})) \\
&= \text{var}(K\widehat{\mu}^{\mathcal{H}}) + \mathbb{E}(l^{-1})\widehat{\Sigma}^{\mathcal{H}} \\
&= \text{diag}(\widehat{\mu}^{\mathcal{H}}) \, \text{var}(k) \, \text{diag}(\widehat{\mu}^{\mathcal{H}}) + \frac{b_l}{a_l - 1}\widehat{\Sigma}^{\mathcal{H}}.
\end{aligned}$$

Using the same equality, we can evaluate the variance of k as

$$\mathrm{var}(k) = \mathrm{var}(\mathbb{E}(k|q,r^{-1})) + \mathbb{E}(\mathrm{var}(k|q,r^{-1}))$$

$$= \mathrm{var}(q \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}) + \mathbb{E}(r^{-1})I_d$$

$$= \sigma_q^2 \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} + \frac{b_r}{a_r - 1}I_d.$$

We obtain

$$\mathrm{var}(\theta) = \mathrm{diag}(\widehat{\mu}^{\mathcal{H}}) \left(\sigma_q^2 \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} + \frac{b_r}{a_r - 1}I_d \right) \mathrm{diag}(\widehat{\mu}^{\mathcal{H}}) + \frac{b_l}{a_l - 1}\widehat{\Sigma}^{\mathcal{H}}$$

$$= \frac{b_l}{a_l - 1}\widehat{\Sigma}^{\mathcal{H}} + \frac{b_r}{a_r - 1}\mathrm{diag}(\widehat{\mu}^{\mathcal{H}})^2 + \sigma_q^2 \widehat{\mu}^{\mathcal{H}t}\widehat{\mu}^{\mathcal{H}}.$$

and the element of the correlation matrix in i th row and j th column ($i \neq j$) is thus

$$\frac{\frac{b_l}{a_l-1}\widehat{\Sigma}^{\mathcal{H}}(i,j) + \sigma_q^2 \widehat{\mu}^{\mathcal{H}}(i)\widehat{\mu}^{\mathcal{H}}(j)}{\left(\frac{b_l}{a_l-1}\widehat{\Sigma}^{\mathcal{H}}(i,i) + (\frac{b_r}{a_r-1} + \sigma_q^2)\widehat{\mu}^{\mathcal{H}}(i)^2 \right)^{1/2} \left(\frac{b_l}{a_l-1}\widehat{\Sigma}^{\mathcal{H}}(j,j) + (\frac{b_r}{a_r-1} + \sigma_q^2)\widehat{\mu}^{\mathcal{H}}(j)^2 \right)^{1/2}}.$$

Under suitable conditions of regularity, $\widehat{\mu}^{\mathcal{H}}$ converges to $\theta_{\mathcal{H}}$ and $\frac{1}{N_{\mathcal{H}}}\widehat{\Sigma}^{\mathcal{H}}$ converges to the Fisher information $I(\theta_{\mathcal{H}})$ as $N_{\mathcal{H}}$ tends to infinity, which ends the proof of the first assertion. The second assertion is trivial when you recall that $\mathbb{E}(r^{-1}) = \frac{b_r}{a_r-1}$. $\qquad\square$

3 Basic examples

In this Section, we describe how to apply our methodology to three different and common examples. We first deal with both a regression model and an auto-regressive model, before addressing a hierarchical Poisson model.

3.1 Regression model

We will consider a Gaussian regression model i.e. for each $t = 1, \ldots, n$

$$y_t = f_t(\theta) + \epsilon_t \tag{5}$$

where $\epsilon_1, \ldots, \epsilon_N$ are assumed independent and identically distributed with common distribution $\mathcal{N}(0, \sigma^2)$.

We will use the notation $f(\theta) = (f_1(\theta), \ldots, f_N(\theta))$ for short, where θ designate the parameters of interest. With these notations, since $\eta = (\theta, \sigma^2)$, the likelihood of the model considered is

$$L(y|\eta) = \left(\sqrt{2\pi}\sigma \right)^{-N} \exp \left(-\frac{1}{2}\|y - f(\theta)\|^2 \right). \tag{6}$$

For σ^2 we use a non-informative prior (we chose $\pi(\sigma^2) \propto \sigma^{-2}$) 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 parameter θ we use the informative hierarchical prior as described in Section 2.2

On the sample $y^{\mathcal{H}}$ we use the following non-informative prior

$$\pi_{\mathcal{H}}(\eta_{\mathcal{H}}) = \pi_{\mathcal{H}}(\theta_{\mathcal{H}}, \sigma_{\mathcal{H}}^2) \propto \sigma_{\mathcal{H}}^{-2}. \tag{7}$$

This prior is non-informative in the sense that it matches Jeffreys' prior distribution on $\sigma_{\mathcal{H}}^2$ for a Gaussian linear regression (see Marin and Robert, 2014, on page 73). It leads to the following posterior distribution

$$\pi_{\mathcal{H}}(\eta_{\mathcal{H}}|y) \propto L(y^{\mathcal{H}}|\eta_{\mathcal{H}})\pi_{\mathcal{H}}(\eta_{\mathcal{H}}) \tag{8}$$
$$\propto \sigma_{\mathcal{H}}^{-N_{\mathcal{H}}-2} \exp\left(-\frac{1}{2}\sigma_{\mathcal{H}}^{-2}\|y - f(\theta_{\mathcal{H}})\|_2^2\right).$$

We consider a polynomial regression model of order p ($p \geq 0$) on the interval $[-1, 1]$, meaning that observations from this model can be written for each $t = 1, \ldots, n$

$$y_t = \sum_{i=0}^{p} \theta_i x_t^p + \epsilon_n \tag{9}$$

where ϵ_n are i.i.d. with Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and where $x_1, \ldots, x_n$ are equally spaced over $[-1, 1]$.

Here again, we only show some of the results we obtained in various settings. Our goal here is to show, for a fixed sample size, the robustness of our method to non-exact similarity between the two datasets

First situation We simulate $N_{\mathcal{H}} = 200$ observations $y^{\mathcal{H}}$ from model (9) with $p = 4$ and coefficients
$$\theta^{\mathcal{H}} = (2, -1, 3, 1, 2)$$

and $N = 10$ observations y with coefficients

$$\theta = \rho\theta^{\mathcal{H}},$$

where ρ will be chosen later, using $\sigma^2 = 1$ for the variance of the noise for $y^{\mathcal{H}}$ and $\sigma^2 = 4$ for y. We then estimate the coefficients θ from the small sample y available using the Gaussian informative prior we presented above. To assess the performance of our method, we compare and aggregate the results over 100 repetitions (sampling new observations for each repetition) with those obtained from a non-informative vague prior and those coming from MLE estimations.

Second situation The second situation is different from the first in that θ is not proportional to $\theta^{\mathcal{H}}$ anymore. Here we chose

$$\theta_0 = \rho\theta_0^{\mathcal{H}}$$
$$\theta_1 = \rho\theta_1^{\mathcal{H}}$$
$$\theta_2 = \theta_2^{\mathcal{H}}$$
$$\theta_3 = \theta_3^{\mathcal{H}}$$
$$\theta_4 = \theta_4^{\mathcal{H}}$$

and the rest of the setup is identical.

Results. Figure 1 shows the errors of estimation for the four coefficients i.e. the true values minus the estimated values (the posterior mean for Bayesian models) when ρ varies from 1 to 0.5 for both situations and Figure 2 displays the actual estimated polynomials for both situations when $\rho = 0.5$.

Once again the benefits of using the informative prior over the non-informative prior stand out. Whether the similarity between the two parameters θ and $\theta^{\mathcal{H}}$ is based upon proportionality or not the informative prior helps the estimation of all the coefficients in a very effective way.

As the sample size N decreases from 100 to 10, the MLE and the non-informative Bayesian method tend to perform rather poorly whereas the use of the informative Bayesian prior clearly benefits the estimation of the coefficients in this situation where observations $y^{\mathcal{H}}$ and y come from exactly the same model.

3.2 Auto-regressive model

We consider an AR(p) model ($p > 0$) with coefficients θ, and variance of noise σ^2, i.e. denoting y_n its observation at time n, for each n

$$y_n = \sum_{i=1}^{p} \theta_i y_{n-i} + \epsilon_n \tag{10}$$

where ϵ_n are i.i.d. with Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

As in the previous example, we choose for θ the informative hierarchical prior described in Section 2.2, and for σ^2 we take the non-informative prior proportional to σ^{-2}. On the historical sample $y^{\mathcal{H}}$, we again choose the non informative prior (7) (see Marin and Robert, 2014, on page 219).

We tested our method on many different such models (with varying orders, coefficients, and variance of noise) and present below only the results obtained for an auto-regressive model of order 4 ($p = 4$) and show the effect of the sample size N.

We simulate $N_{\mathcal{H}} = 400$ observations $y^{\mathcal{H}}$ from model (10) with coefficients

$$\theta^{\mathcal{H}} = (1.7, -0.23, -0.833, 0.3528)$$

(the corresponding polynomial having roots $-0.7, 0.7, 0.8$ and 0.9), and also simulate $N = 100, 30, 10$ observations y from the same AR(4) model with coefficients

$$\theta = \theta^{\mathcal{H}},$$

using in each case $\sigma^2 = 1$ for the variance of the noise. We then estimate the coefficients θ from the observations y available using the Bayesian prior we presented above. To assess the performance of our method, we compare and aggregate the results over 100 repetitions (sampling new observations for each repetition) with those obtained from a non-informative vague prior and those coming from MLE estimations. Figure 3 shows the errors of estimation for the four coefficients i.e. the true values minus the estimated values (the posterior mean for Bayesian models).

3.3 Hierarchical Poisson model

We consider $y_1, \ldots, y_N$ independent random variables such that for each j, y_j follows a Poisson distribution with parameter λ_j. We assign a Gamma distribution for the parameters λ_i :

$$y_i \sim \mathcal{P}(\lambda_i)$$
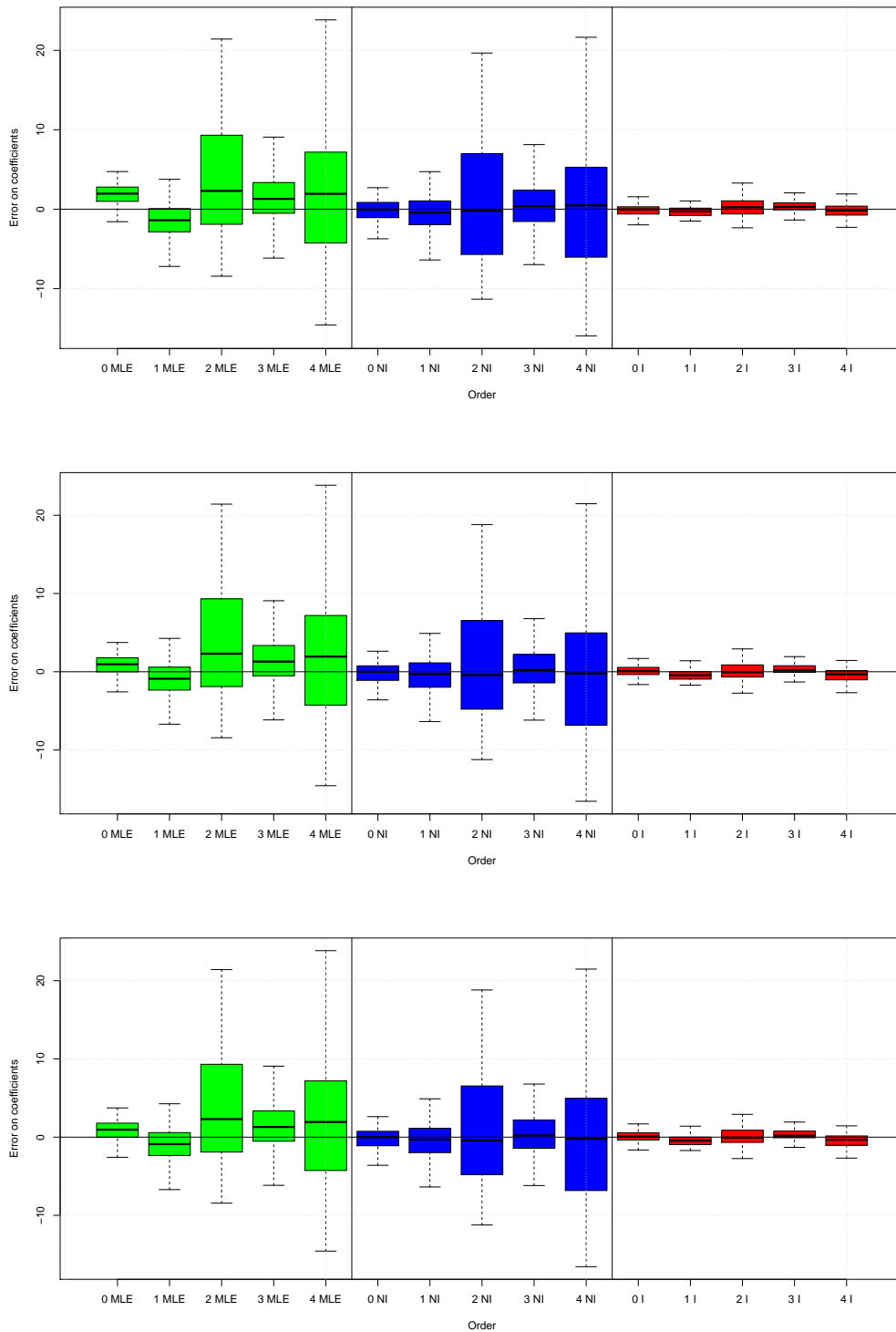$$\lambda_i \sim \Gamma(a, b)$$

Figure 1: Estimation errors on the four coefficients of the polynomial regression model (9), using MLE (in green), non-informative Bayesian (NI, in blue) and informative Bayesian (I, in red) methods, when $\rho = 1$ (top row), $\rho = 0.5$ in the first situation (middle row), $\rho = 0.5$ in the second situation (bottom row)
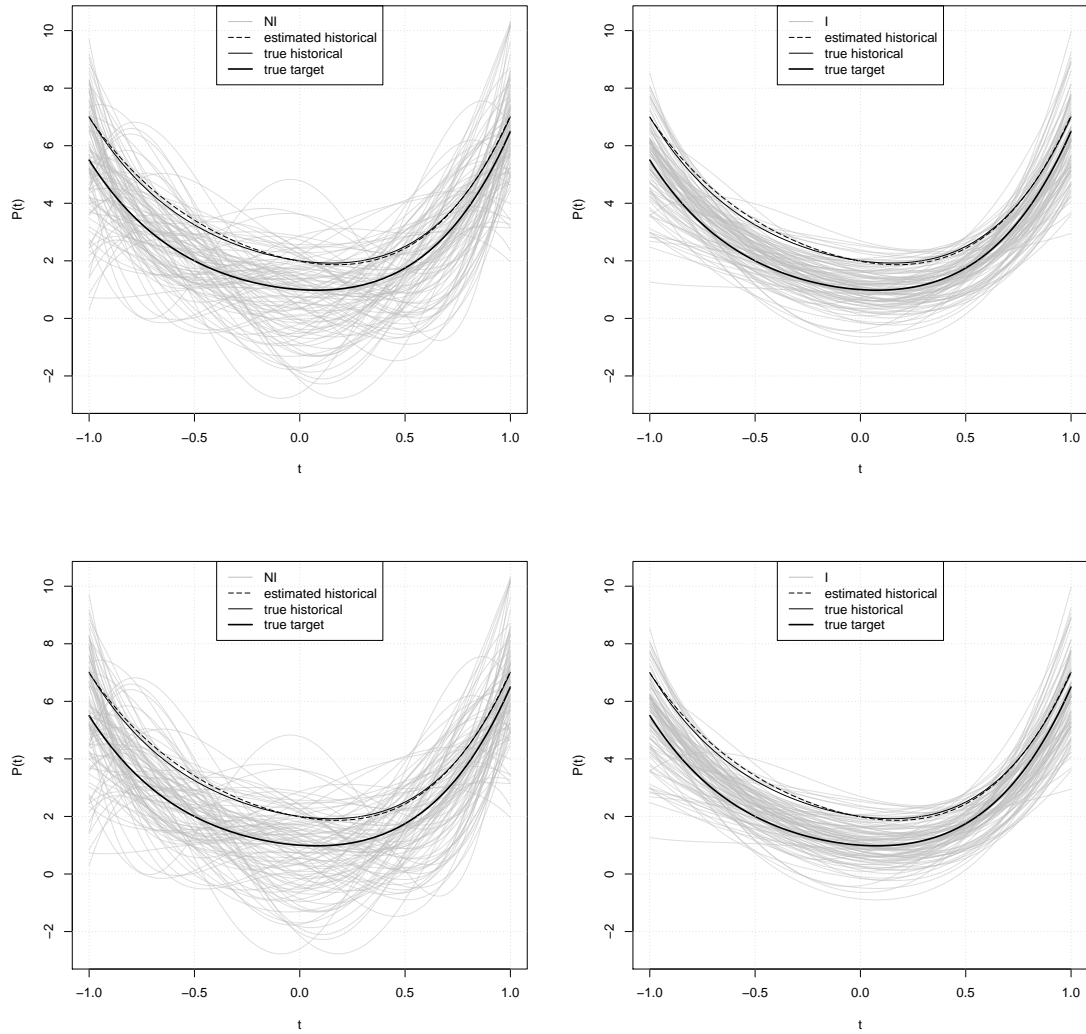
Figure 2: Estimated polynomials using non-informative Bayesian prior (left) and informative Bayesian (right). The thin solid line represents the true polynomial corresponding to $\theta^{\mathcal{H}}$, the dashed line the estimated polynomial on observations $y^{\mathcal{H}}$, and the thick solid line represents the true polynomial corresponding to θ. Top row shows results for the first situation, while bottom row shows results for the second situation.
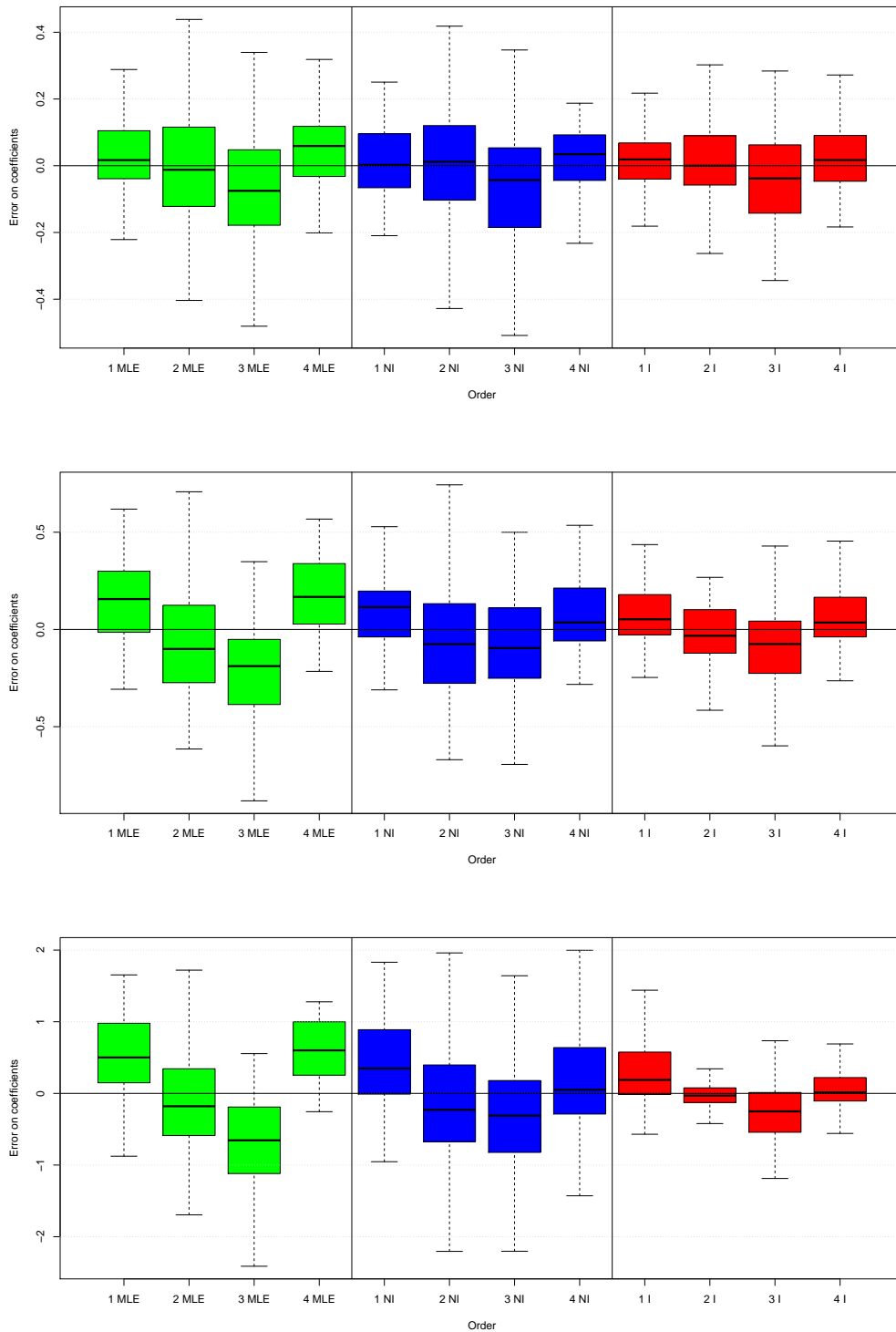
Figure 3: Estimation errors on the four coefficients of the auto-regressive model (10), using MLE (in green), non-informative Bayesian (NI, in blue) and informative Bayesian (I, in red) methods, for three different sample sizes : $N = 100$ (top row), $N = 30$ (middle row), $N = 10$ (bottom row)

When a large sample is available, we can choose a non informative distribution on the parameter $\theta = (a, b)$, for example the Jeffreys prior which is given by

$$\pi_J(a, b) = \frac{1}{b} \sqrt{a(\log(\Gamma(a))'' - 1}$$

(see Yang and Berger (1998)). When a goes to zero or to infinity, $\sqrt{a(\log(\Gamma(a))'' - 1}$ is equivalent to $\frac{1}{\sqrt{a}}$ (see Abramovitz and Stegun (1965) chapter 6). To simplify the computational aspects, we consider a prior distribution which behaves as the Jeffreys prior when a tends to zero and to infinity. We assume that a and b are independent, the marginal distributions of a and b are Gamma distributions with parameters $(1/2, \epsilon)$ and (ϵ, ϵ) respectively. The parameter ϵ is fixed small enough.

For this choice of prior, Figure 4 illustrates that the Bayes estimates do not perform well for the small sample.

In the context of Section 2, there is a large sample $y^{\mathcal{H}}$ sharing some similarities with the small sample of interest y. Therefore we can construct the prior distribution on (a, b) using the sample $y^{\mathcal{H}}$ based on either of the two following strategies

1. we assign on (a, b) the hierarchical prior defined in (4) with $\theta = (a, b)$.

2. the parameter (a, b) are fixed, we take the values of the estimates calculated on the large sample $y^{\mathcal{H}}$ (See Albert (2009); Gelman et al. (2013))

Figure 4 shows the improvement brought by the informative hierarchical model upon non informative model. On independent replications, we estimate the mean square error on the estimation of θ_i

$$E_N = \frac{1}{N} \sum_{i=1}^{N} |\lambda_i - \mathbb{E}(\lambda_i | y)|^2. \tag{11}$$

The results are given in Figure 5. When the two populations have exactly the same parameters, it is not surprising to obtain better results using fixed parameters. However the simulation shows also that this strategy is not robust and leads to worse results than the non informative prior when the true parameters differ. The hierarchical prior on (a, b) gives some very satisfying results in terms of quality of the estimation and robustness to variation of the parameters between the sample $y^{\mathcal{H}}$ and y.
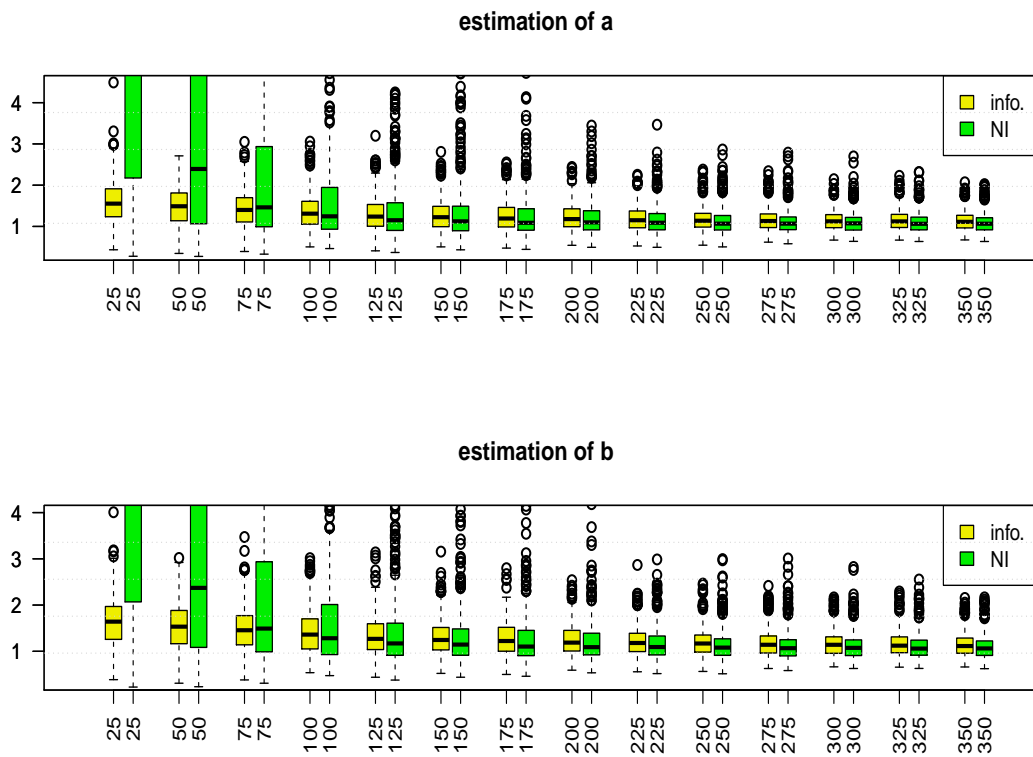
estimation of a

estimation of b

Figure 4: Estimation of the hyperparameters $\theta = (a, b)$ using hierarchical informative prior (Info.) and non informative prior as function of sample sizes N. The true gamma parameters are $(a_{\mathcal{H}}, b_{\mathcal{H}}) = (a, b) = (1, 1)$. The boxplot is constructed from 500 independent replications.
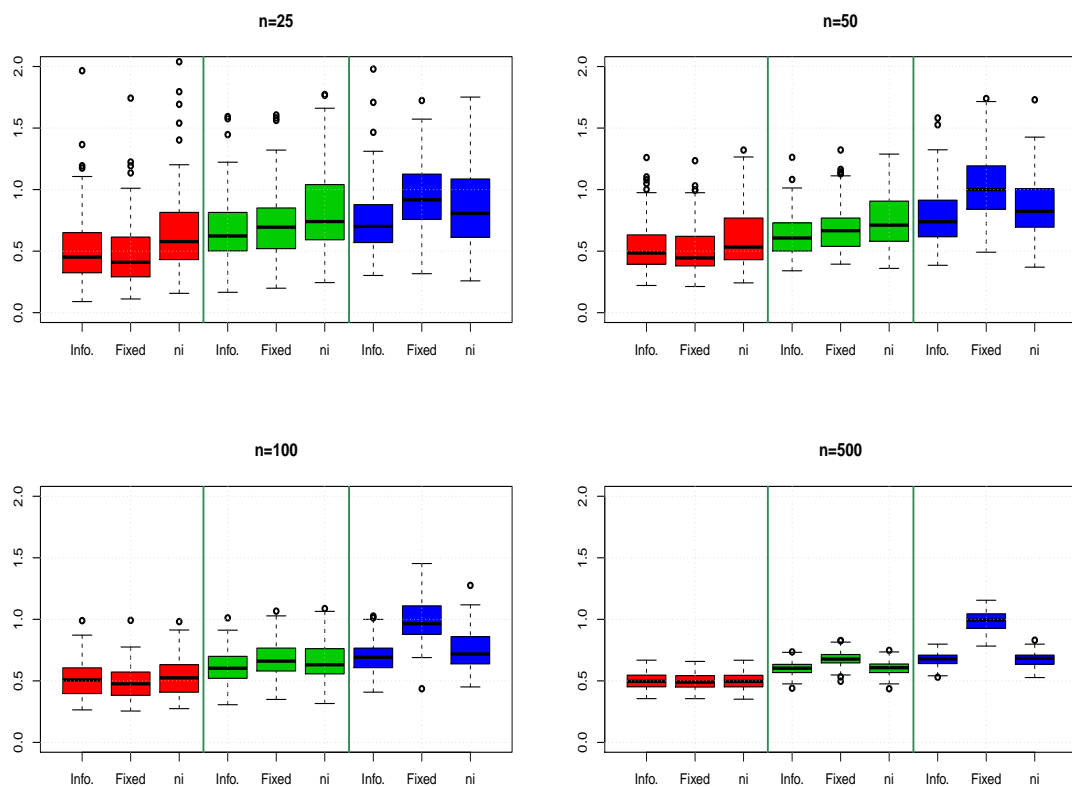
Figure 5: Mean square error E_n defined in (11) for the three choice of prior : hierarchical prior (Info.), fixed hyperparameter (Fixed) and non-informative prior (ni). The parameters are $\theta = (a, b)$: $(1, 1)$ (red), $(9/4, 3/2)$ (green), $(4, 2)$ (blue). The sample sizes are $N = 25$, 50, 100, 500 and $N_{\mathcal{H}} = 5000$. The gamma parameters of $y^{\mathcal{H}}$ are $\theta_{\mathcal{H}} = (a_{\mathcal{H}}, b_{\mathcal{H}}) = (1, 1)$. The boxplot is constructed from 500 independent replications

14

4 Electricity load forecasting

Modelling and forecasting the electricity load 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: 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 (see Cottet and Smith, 2003, for example)).

Electricity load usually involves a large predictable component due to its very strong daily, weekly and yearly periodic behaviour and the exogenous variables most commonly used to forecast the electricity load are weather-based (see Al-Zayer and Al-Ibrahim, 1996; Taylor and Buizza, 2003, for an example). For the French electricity load specifically, the importance of the temperature and cloud-cover is underlined in Menage et al. (1988) and Bruhns et al. (2005). Engle et al. (1986); Ramanathan et al. (1997) also present other weather-related models.

Some authors work with univariate time series models: Taylor (2003) builds a double seasonal exponential smoothing for the British electricity load, and Cugliari (2011) treats the load curve as a functional-valued autoregressive Hilbertian process to use a non-parametric approach relying on the wavelet transform. Other authors (see Soares and Medeiros, 2008; Ramanathan et al., 1997; Dordonnat et al., 2008, for example) choose multiple-equation models where the various hours of the day share the same equation, while the associated parameters differ.

Bayesian methods have also been used in the past to solve electricity load forecasting problems. Amongst them, Minka (1999) deals with Bayesian linear regression and Harrison and Stevens (1976) focus on dynamic linear models from the Kalman filter point of view and Smith (2000) choose a Bayesian semi-parametric regression model. To the best of our knowledge though, no previous attempt has been made in the literature to help improve the forecasting of an intraday electricity load model on a short dataset, based upon another dataset using a hierarchical Bayesian model.

4.1 Regression model

In the next few paragraphs, we specify the regression model that we use to model the electricity load.

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 and is usually formulated as follows: for $t = 1, \ldots, N$,

$$y_t = x_t^{(1)} x_t^{(2)} + x_t^{(3)} + \epsilon_t \tag{12}$$

$$x_t^{(1)} = \sum_{j=1}^{d_{11}} \left[z_j^{\cos} \cos \left(\frac{2j\pi}{365.25} t \right) + z_j^{\sin} \sin \left(\frac{2j\pi}{365.25} t \right) \right] + \sum_{j=1}^{d_{12}} \omega_j \mathbb{1}_{\Omega_j}(t),$$

$$x_t^{(2)} = \sum_{j=1}^{d_2} \psi_j \mathbb{1}_{\Psi_j}(t),$$

$$x_t^{(3)} = g(T_t - u) \mathbb{1}_{[T_t, +\infty[}(u),$$

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
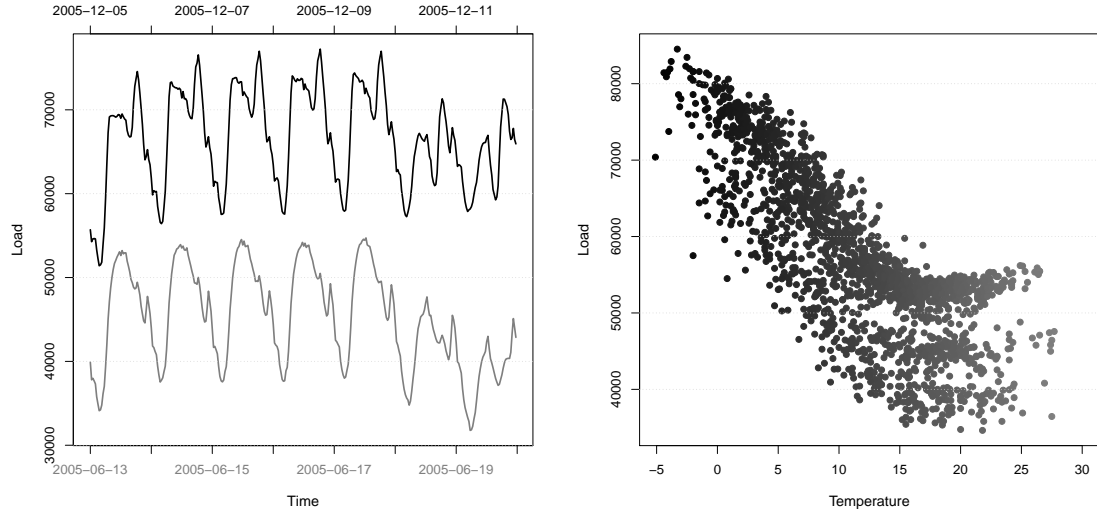
Figure 6: Left : French Electricity load from 13/06/2005 to 29/06/2005 (in grey) and from 05/12/2005 to 11/12/2005 (in black). The load is expressed in MW. Notice the daily patterns of the electricity load are not the same during summer and winter. Right : French Electricity load at 10:00 over 5 years against temperatures. The load seems to increase linearly with the temperature below a certain threshold.

electricity consumption behaviour. The left part of Figure 6 shows the typical behaviour of the electricity load 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 $(\Psi_j)_{j \in \{1,...,d_2\}}$ 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 6 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. Note that the threshold is sought within the range of the observed temperatures, i.e. $u \in [\underline{u}, \overline{u}]$ with

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

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 6.

4.2 Priors

For the sake of both clarity and completeness, we present the full Bayesian model in use for all our applications in the lines below. Let us denote here again $\eta = (\theta, \sigma^2)$ where θ designates the parameters of interest in the electricity load model (i.e. $z_j, \omega_j, \psi_j, g,$ and u).

We opted for the Gaussian strategy (see Section 2.2) for the regression model as was described in Section 3.1, meaning that the non-informative prior distribution $\pi_{\mathcal{H}}$ used with the historical sample $y^{\mathcal{H}}$ is

$$\pi_{\mathcal{H}}(\eta_{\mathcal{H}}) = \pi_{\mathcal{H}}(\theta_{\mathcal{H}}, \sigma_{\mathcal{H}}^2) \propto \sigma_{\mathcal{H}}^{-2}.$$

The informative prior π used with the small sample y is defined hierarchically and can be summarised as

$$\pi(\theta, \sigma^2, k, l, q, r) \propto \pi(\theta|k, l)\pi(k|q, r)\pi(l)\pi(q)\pi(r)\pi(\sigma^2)$$

with:

$$\theta|k, l \sim \mathcal{N}(K\widehat{\mu}^{\mathcal{H}}, l^{-1}\widehat{\Sigma}^{\mathcal{H}}) \qquad\qquad q \sim \mathcal{N}(1, \sigma_q^2)$$

$$k|q, r \sim \mathcal{N}(q(1, \ldots, 1)', r^{-1}I_d) \qquad\qquad r \sim \mathcal{G}(a_r, b_r)$$

$$l \sim \mathcal{G}(a_l, b_l) \qquad\qquad \pi(\sigma^2) \propto \sigma^{-2},$$

where $K = \mathrm{diag}(k)$, where $\widehat{\mu}^{\mathcal{H}}$ and $\widehat{\Sigma}^{\mathcal{H}}$ are the posterior mean and posterior variance of the parameter $\theta_{\mathcal{H}}$ and 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.

Remark 6. *Using either the non-informative prior distribution or the informative prior distribution with the model* (12) *leads to a proper posterior distribution. The proofs, as well as the MCMC algorithms we designed to derive numerical estimators, are pushed back into the Appendix (see Section 5).*

4.3 Datasets

The historical dataset $y^{\mathcal{H}}$ 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 benefits of choosing our informative prior to predict electricity load on two short datasets : the first dataset y 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 $y^{\mathcal{H}}$ does, but not exactly. The population behind the second dataset y' corresponds to a subset of the population behind $y^{\mathcal{H}}$ and represents around 50% of the total load of $y^{\mathcal{H}}$.

4.4 Benchmark against standard methods

For this application, we only consider the days for which no special tariffs are enforced: the so-called EJP ("Effacement jour de pointe" = peak tariff days) are removed from the dataset beforehand to ensure the signal studied is 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) are also removed from the dataset for this first application, so that we may benchmark our method against others on a smooth and rather easy-going signal, and not put any one method at a disadvantage due to the

	Estimation $y^{\mathcal{H}}$	Estimation y	Prediction y
Case 1	1099	125	28
Case 2	1099	144	38

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

| | 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.

signal's specificity. The temperature considered in the model is the average temperature over France for the period of study, and the cooling threshold is chosen to be 16°C throughout the 48 instants of the day.

We benchmark our Bayesian method with informative prior against four alternative methods, comparing their predictions on dataset y in two configurations. Roughly speaking, for our first experiment (Case 1) we estimate the model for y over the period ranging from 12/01/2009 to 06/30/2010 and predict the next 30 days (same as the application presented Section 4.5), while for our second experiment (Case 2) we estimate the model for y over the period ranging from 01/01/10 to 07/31/10 and predict the previous 30 days. We expect Case 1 to be the easy case and Case 2 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 benchmark against our own Bayesian informative method, rely on four different techniques: the first one is the Bayesian non-informative method that we presented earlier in Section 2.2 (recall that it is meant to be an equivalent to the maximum likelihood approach), the second involvedsnon-parametric estimation with kernels (see Fan and Yao, 2005), the third is a double exponential smoothing (see Taylor, 2003) and the fourth and last is an ARIMA model. Note that for the second experiment the data available obviously has to be time-reversed in order to apply the last three alternatives methods, since time-dependence plays an important role for them. The ARIMA model is automatically selected (see Hyndman and Khandakar, 2008) as the best model with regard to the AIC criterion.

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. 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.

4.5 Role of the hyperparameters

For this application, the setup is the same as the one described at the start of Section 4.4. Our aim here is 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 4.4, they also provide a measure of similarity between the short datasets of interest and the dataset used to build the prior. We estimate the model for y and y' over the period ranging

from 12/01/2009 to 06/30/2010 and predict the next 30 days.

Figure 7 shows the predictive quality of the model for both populations y and y' using the informative prior as well as the posterior densities of the similarity coefficients k_j at midday (we also checked the other 47 instants but the look of them was nearly identical to the one we present). These densities are much more peaked and also centred closer to 1 for population y than they are for population y'. It thus seems to indicate that dataset y is more similar to $y^{\mathcal{H}}$ than y' is, confirming our prior knowledge that $y^{\mathcal{H}}$ and y cover approximately the same population whereas y' represents around 50% of $y^{\mathcal{H}}$: this value of 50% is also visible on Figure 7 where we observe two densities centred around 0.5, which correspond to the similarity coefficients between the offsets ω_j of $y^{\mathcal{H}}$ and these of y'.

Figure 8 displays the boxplots for the posterior densities of q and $1/\sqrt{r}$ and seem to corroborate the fact that y is more similar to $y^{\mathcal{H}}$ than y' 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 y than y'. The estimated mean and variance of $1/\sqrt{r}$ are also smaller on y than y'.

The estimated values of q and r provide some information about the similarity between the datasets considered. Notice also that the best predictive performance is obtained when the similarity between the two datasets is strongest.

4.6 Role of the sample size

For this application, the setup is almost identical to the one described at the start of Section 4.4 but the temperature used in the model is not the average temperature over France anymore but a transformation of it: it is 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 is fixed at 18°C throughout the 48 instants of the day, and this time, the summer holiday break is not removed from the dataset (but the winter holiday break and the bank holidays still are), so that the model could benefit from (and be tested on) the August months in general. Note that for $y^{\mathcal{H}}$ we use the same dataset as we used for our two first applications.

We put the focus on the length of the estimation period on y while keeping the same prediction window. We successively choose 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 y from 12 months to only 6 months, removing 2 months at a time. We then predict the next 6 upcoming months i.e. from 01/01/2011 to 06/30/2011. The diagram in Figure 9 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 10 shows the average error in prediction for each month. It is important to note that the use of the non-informative prior leads to overfitting the model: the results presented in Table 3 show that as the estimation window goes smaller, the estimation error decreases while the prediction error grows very quickly. A close inspection of the posterior densities of the different parameters of the model reveals that the bias induced by the increasing lack of data is mainly seasonal: this is due to the seasonality coefficients of the model being overfit. Choosing the informative prior over the non-informative prior makes the estimation and prediction of the
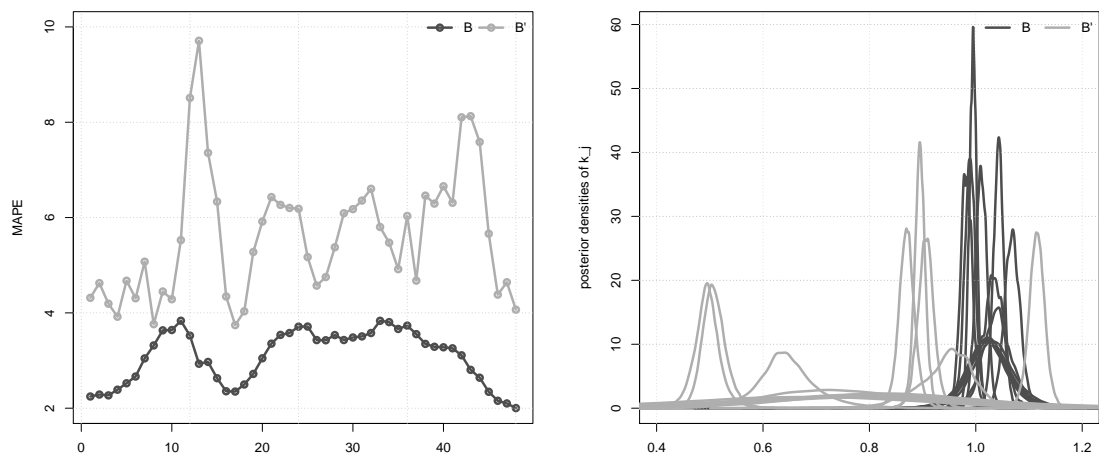
Figure 7: Quality of the predictions (MAPE in %) averaged for each instant, for populations y and y' (left). Posterior densities of the similarity coefficients k_j for populations y and y' at midday (right).
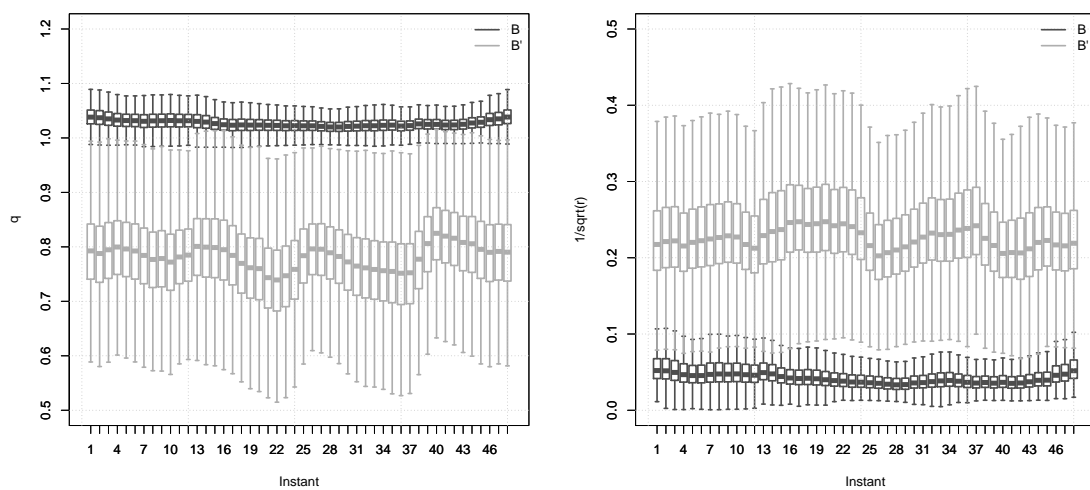

Figure 8: Boxplots of the posterior densities of q (left) and $1/\sqrt{r}$ (right) (mean and standard deviation of the similarity coefficients k_j) for populations y and y', throughout the 48 instants of the day.
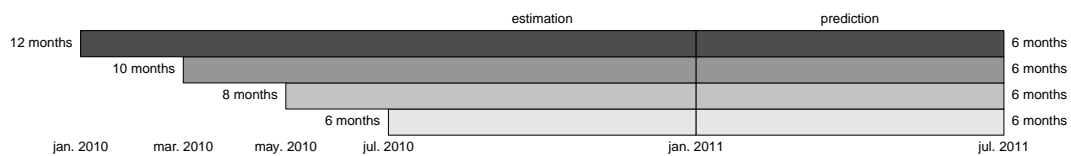

Figure 9: Ranges of the estimation (from 12 to 6 months) and prediction (6 months) time-windows for the 4 scenarios considered.

		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).

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 are 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.
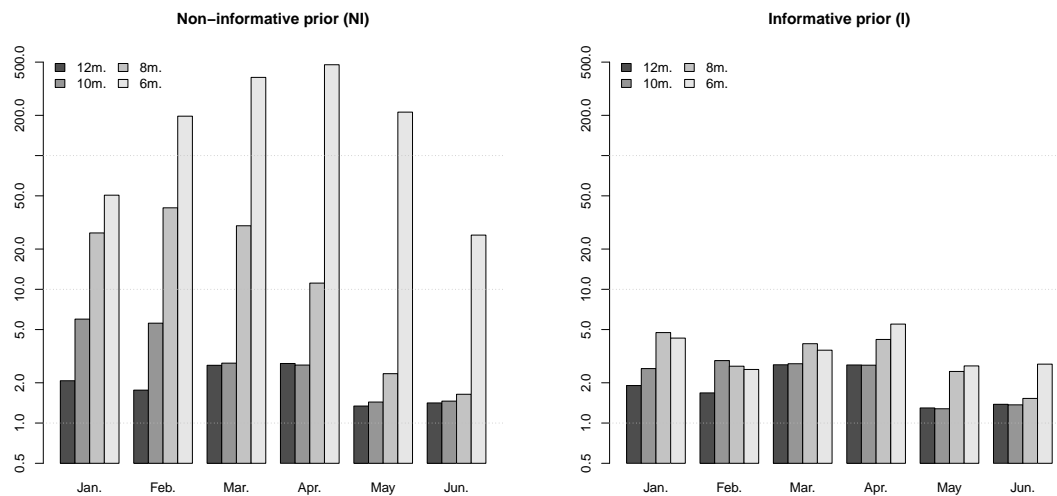

Figure 10: Quality of the predictions (MAPE in %) averaged for each month (all the instants of the 30 or so days within each month are used to compute these averages), with an estimation period ranging from 12 to 6 months using the non-informative prior (left) and using the informative prior (right). The ordinate axis is in log-scale. Each shade of grey corresponds to a different scenario (as depicted on Figure 9).

5 Appendix

5.1 Notations

Using the notation $M_{i\bullet}$ for the i-th row of a matrix M, the non-linear model described in (12) can be re-written in the following condensed way, as a particular case of the regression model (5): for $t = 1, \dots, N$,

$$y_t = (A_{t\bullet}\alpha)(B_{t\bullet}\beta + C_t) + \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u) + \epsilon_t. \tag{13}$$

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

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

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.

5.2 Technical proofs

Proposition 7. *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 (13) and the posterior measure is then a well-defined (proper) probability distribution.

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

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

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

Proposition 8. *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 (13), the posterior measure (8) is then a well-defined (proper) probability distribution.

Proof. Notice first that

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

and observe then that

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

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

$$\|y - f((\alpha, \beta_0, \gamma, u_0))\|_2^2 = \sum_{t=1}^N \left[y_t - (B_{t\bullet}\beta_0 + C_t)A_{t\bullet}\alpha - (T_t - u_0)\mathbb{1}_{[T_t, +\infty[}(u_0)\gamma \right]^2$$

$$= \|y - A_*(\beta_0, u_0)\alpha_*\|_2^2,$$

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

$$\|y - f(\theta)\|_2^{-N} \sim \|y - A_*(\beta_0, u_0)\alpha_*\|_2^{-N}. \tag{14}$$

The triangular inequality applied to the right hand side of (14) gives

$$\|y - A_*(\beta_0, u_0)\alpha_*\|_2^{-N} \le \left| \|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2 \right|^{-N}. \tag{15}$$

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

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

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

$$\left| \|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2 \right|^{-N} \sim \lambda^{-N/2}\|\alpha_*\|_2^{-N}. \tag{16}$$

Combining (14), (15) and (16) together, we see that the integrability of the left hand side of (14) 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 immediate for $N > d_\alpha + 1$, as can be seen via a Cartesian to hyperspherical re-parametrisation.

The previous paragraph thus ensures the integrability of $\|y - f(\theta)\|_2^{-N}$ 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 $\|y - f(\theta)\|_2^{-N}$ over $\{(\beta, u) \in B_+^{d_\beta}(0,1), \|\alpha_*\| \in]M, +\infty[\}$.

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

Remark 9. *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".*

5.3 MCMC algorithms for the estimation of the posterior density

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 Robert and Casella, 2009, page 230, for a 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.

5.3.1 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 $(\eta_1, \ldots, \eta_M)$ according to the posterior distribution $\pi(\eta|y)$ corresponding to the informative prior we presented earlier. Denoting $\mathcal{N}(\mu, \Sigma, S)$ the S-truncated Gaussian distribution with mean μ and covariance Σ, the algorithm goes as follow:

Step 1. Initialise η_1 such that $\pi(\eta_1|y) \neq 0$

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

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

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

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

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

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

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

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

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

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

$$k_{t+1} \sim \mathcal{N}\left(\mu_{t+1}^k, \Sigma_{t+1}^k\right)$$

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

$$\gamma_{t+1} \sim \mathcal{N}\left(\mu_{t+1}^g, \Sigma_{t+1}^g\right)$$

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

$$\beta_{t+1} \sim \mathcal{N}\left(\mu_{t+1}^b, \Sigma_{t+1}^b, B_+^{d_\beta}(0,1)\right)$$

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

$$\alpha_{t+1} \sim \mathcal{N}\left(\mu_{t+1}^a, \Sigma_{t+1}^a\right)$$

(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 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 $\mu^\alpha_{t+1}, \Sigma^\alpha_{t+1}, \mu^\beta_{t+1}, \Sigma^\beta_{t+1}, \mu^\gamma_{t+1}, \Sigma^\gamma_{t+1}, \mu^k_{t+1}$ and Σ^k_{t+1}, are to be found in Launay (2012).

5.3.2 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 $(\eta_1, \ldots, \eta_M)$ according to the posterior distribution $\pi(\eta|y)$ corresponding to the non-informative prior we presented earlier. Denoting $\mathcal{N}(\mu, \Sigma, S)$ the S-truncated Gaussian distribution with mean μ and covariance Σ, the algorithm goes as follows:

Step 1. Initialise η_1 such that $\pi(\eta_1|y) \neq 0$

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 $\delta_t \sim \mathcal{N}(0, \Sigma_{\mathrm{MH}})$, 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

25

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 $\mu_{t+1}^{\alpha}, \Sigma_{t+1}^{\alpha}, \mu_{t+1}^{\beta}, \Sigma_{t+1}^{\beta}, \mu_{t+1}^{\gamma}$, and Σ_{t+1}^{γ}, are to be found in Launay (2012).

Acknowledgments

References

Abramovitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York, Dover Publications.

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.

Albert, J. (2009). *Bayesian computation with R*. Springer, Dordrecht.

Bouveyron, C. and Jacques, J. (2013). Adaptive mixtures of regressions: Improving predictive inference when population has changed. *Communications in Statistics - Simulation and Computation, to appear*.

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*. John Wiley, New York.

Congdon, P. (2010). *Applied Bayesian Hierarchical Methods*. Applied Bayesian Hierarchical Methods, Chapman & Hall, CRC.

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). *Prévision non paramétrique de processus à valeurs fonctionnelles, application à la consommation d'électricité*. PhD thesis, Univérsité Paris Sud XI.

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.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Institute of Mathematical Statistics Monographs)*. Cambridge University Press.

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., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. Chapman & Hall CRC Texts in Statistical Science.

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

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis, Theory and Methods*. Springer.

Harrison, P. and Stevens, C. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society*, 38(3):205–247.

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

Launay, T. (2012). *Bayesian methods for electricity load forecasting*. PhD thesis, Université de Nantes.

Launay, T., Philippe, A., and Lamarche, S. (2012). Consistency of the posterior distribution and MLE for piecewise linear regression. *Electron. J. Statist.*, 6:1307–1357.

Marin, J.-M. and Robert, C. (2014). *Bayesian Essentials with R*. Texts in Statistics. Springer, 2nd edition edition.

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.

Minka, T. P. (1999). Bayesian linear regression. Technical report, 3594 Security Ticket Control.

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.

Robert, C. P. and Casella, G. (2009). *Introducing Monte Carlo Methods with R*. Springer Verlag, 1st edition.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367.

Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear Regression (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

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

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. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8):799–805.

Taylor, J. W. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19(1):57–70.

Yang, R. and Berger, J. (1998). A catalog of noninformative priors. Technical report, Institute of Statistics and Decision Sciences, Duke University.