L M J L

UMR 6629 - Nantes

UNIVERSITÉ DE NANTES

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

Anne Philippe
Laboratoire de Mathématiques Jean Leray
Université de Nantes

Workshop EDF-INRIA, 5 avril 2012, IHP - Paris

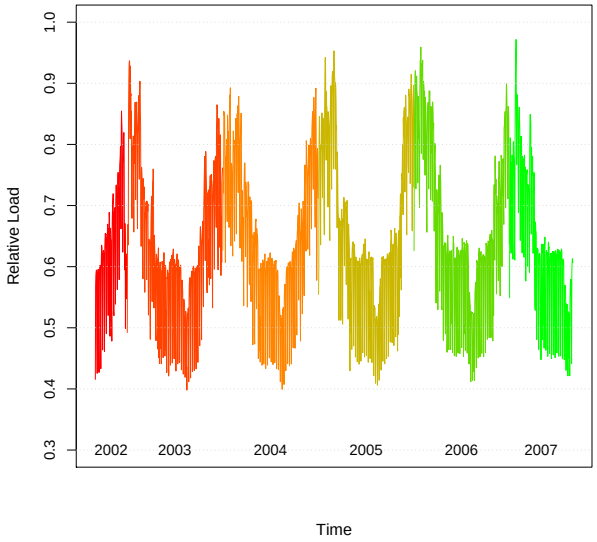
Work in collaboration with
Tristan Launay – LMJL / EDF OSIRIS, R39
and
Sophie Lamarche – EDF OSIRIS, R39

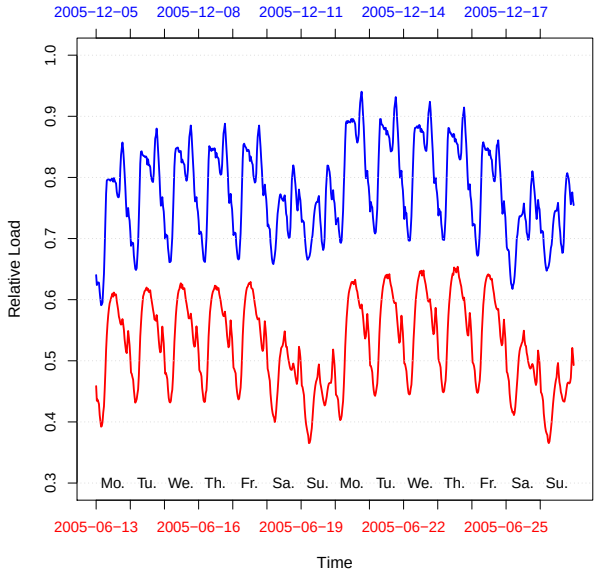
anne.philippe@univ-nantes.fr

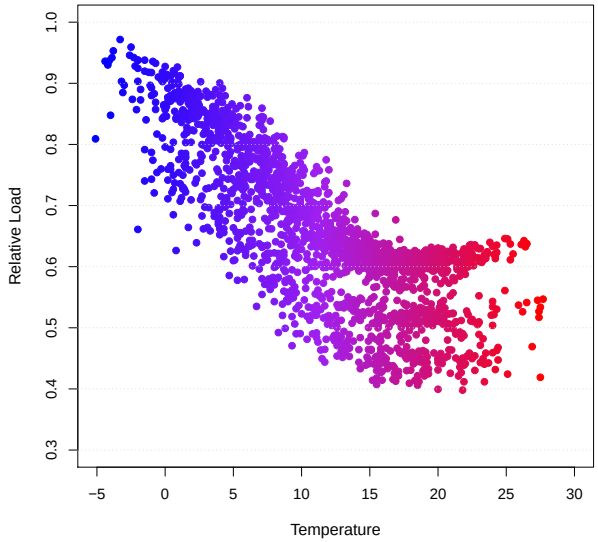
Plan

Plan

The electricity load signal







Comments

- the electricity demand exhibits multiple strong seasonalities : yearly, weekly and daily.
- daylight saving times, holidays and bank holidays induce several breaks
- non-linear link with the temperature.

The model

The electricity load X_t is written as

$$X_t = s_t + w_t + \epsilon_t$$

- s_t : seasonal part
- w_t : non-linear weather-depending part (heating)
- $(\epsilon_t)_{t\in\mathbb{N}}$ i.i.d. from $\mathcal{N}(0, \sigma^2)$

Heating part

$$w_t = \begin{cases} g(T_t - u) & \text{if } T_t < u \\ 0 & \text{if } T_t \geq u \end{cases}$$

- g is the heating gradient,
- u is the heating threshold

Model (cont.)

Seasonal part

$$s_t = \left[\sum_{j=1}^{p} \left\{ a_j \cos\left(\frac{2\pi j}{365.25} \cdot t \right) + b_j \sin\left(\frac{2\pi j}{365.25} \cdot t \right) \right\} + \sum_{j=1}^{q} \omega_j \mathbb{1}_{\Omega_j}(t) \right] \cdot \sum_{j=1}^{r} \kappa_j \mathbb{1}_{K_j}(t)$$

- the average seasonal behaviour with a truncated Fourier series
- gaps (parameters $\omega_j \in \mathbb{R}$) represent the average levels over predetermined periods given by a partition $(\Omega_j)_{j \in \{1,\ldots,d_{12}\}}$ of the calendar.
- day-to-day adjustments through shapes (parameters κ_j) that depends on the so-called days' types which are given by a second partition $(K_j)_{j \in \{1,\ldots,d_2\}}$ of the calendar.

Plan

Bayesian principle

Let $X_1, \ldots, X_N$ be the observations.
The likelihood of the model, conditionnally to the temperatures, is :

$$l(X_1, \ldots, X_N|\theta) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[X_t - (s_t + w_t)\right]^2\right\}$$

Let $\pi(\theta)$ be the *prior* distribution on θ

The bayesian inference is based on

- the *posterior* distribution of θ

$$\pi(\theta|X_1, \ldots, X_N) \propto l(X_1, \ldots, X_N|\theta)\pi(\theta)$$

- The Bayes estimate (under quadratic loss) is $\widehat{\theta} = \mathbb{E}[\theta|X_1, \ldots, X_N]$

Predictive distribution

- $$p(X_{N+k}|X_1, \ldots, X_N) = \int_{\Theta} l(X_{N+k}|\theta)\pi(\theta|X_1, \ldots, X_N)\,\mathrm{d}\theta$$

- The optimal prediction (under quadratic loss) is $\widehat{X}_{N+k} = \mathbb{E}[X_{N+k}|X_1, \ldots, X_N]$

Asymptotic results for the heating part

The observations $X_{1:n} = (X_1, \ldots, X_n)$ depend on an exogenous variable $T_{1:n} = (T_1, \ldots, T_n)$ via the model

$$X_i = \mu(\eta, T_i) + \xi_i := g(T_i - u)\mathbb{1}_{[T_i, +\infty[}(u) + \xi_i,$$

for $i = 1, \ldots, n$.

- $(\xi_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. from $\mathcal{N}(0, \sigma^2)$.
- θ_0 will denote the true value of θ.

$$\theta = (g, u, \sigma^2) \in \Theta = \mathbb{R}^* \times]\underline{u}, \overline{u}[\times \mathbb{R}_+^*.$$

Consistency

Assumptions (A).

1. There exists $K \subset \Theta$ a compact subset of the parameter space Θ such that the MLE $\widehat{\theta}_n \in K$ for any n large enough.
2. Let F be a cumulative distribution function which is continuously differentiable and F' does not vanish over $[\underline{u}, \overline{u}]$

$$F_n(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[T_i, +\infty[}(u), \xrightarrow[n \to +\infty]{} F(u)$$

Theorem (Launay et al [2])

Let $\pi(\cdot)$ be a prior distribution on θ, continuous and positive on a neighbourhood of θ_0 and let U be a neighbourhood of θ_0,
then under Assumptions (A), as $n \to +\infty$,

$$\int_U \pi(\theta | X_{1:n}) \, \mathrm{d}\theta \xrightarrow{a.s.} 1.$$

Asymptotic normality

Theorem (Launay et al [2])

Let $\pi(\cdot)$ be a prior distribution on θ, continuous and positive at θ_0, and let $k_0 \in \mathbb{N}$ such that

$$\int_{\Theta} \|\theta\|^{k_0} \pi(\theta)\, \mathrm{d}\theta < +\infty,$$

and denote

$$t = n^{\frac{1}{2}}(\theta - \widehat{\theta}_n),$$

and $\widetilde{\pi}_n(\cdot|X_{1:n})$ the posterior density of t given $X_{1:n}$, then under Assumptions (A), for any $0 \leqslant k \leqslant k_0$, as $n \to +\infty$,

$$\int_{\mathbb{R}^3} \|t\|^k \left| \widetilde{\pi}_n(t|X_{1:n}) - (2\pi)^{-\frac{3}{2}} |I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2} t' I(\theta_0) t} \right| \mathrm{d}t \xrightarrow{\mathbb{P}} 0,$$

where $I(\theta)$ is the asymptotic Fisher Information matrix.

Plan

Objective

Consider the general model $X_t = s_t + w_t + \epsilon_t$

Aim

the estimation and the predictions of the parametric model over a short dataset denoted B

Problem

- The high dimensionality of a model leads to an overfitting situation, and so the errors in prediction are larger
- the parameter of the model

$$\theta = (\eta, \sigma^2),\ \eta = (\underbrace{a_1, \ldots, a_p, b_1, \ldots, b_p}_{Fourier}, \underbrace{\omega_1, \ldots, \omega_k}_{offsets}, \underbrace{\kappa_1, \ldots, \kappa_r}_{shapes}, \underbrace{g, u}_{heating\ part})$$

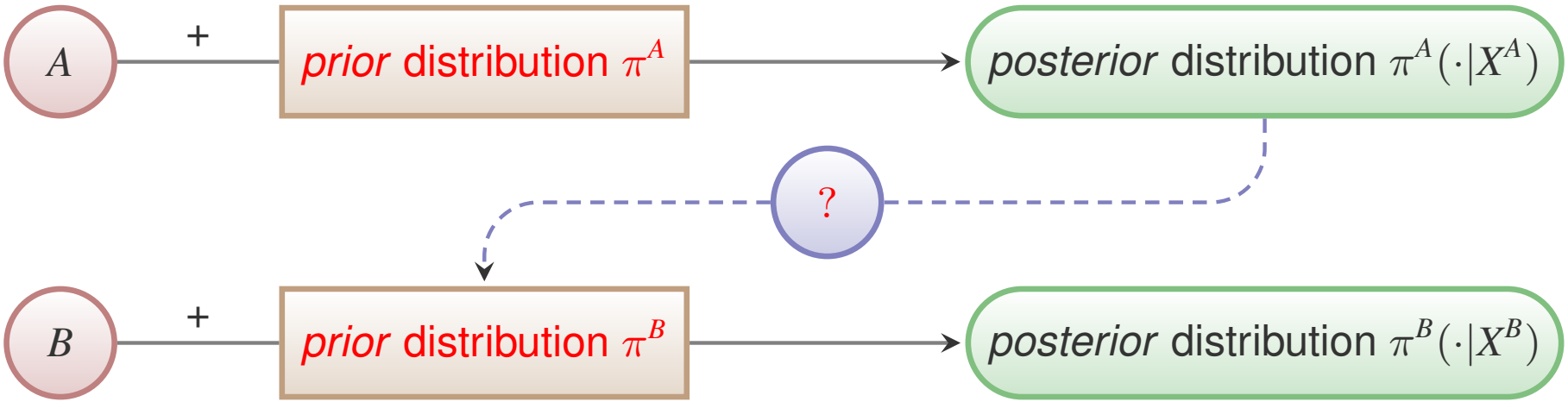
Prior information

- A a long dataset known to share some common features with B

We wish to improve parameter estimations and model predictions over B with the help of A.

Choice of the prior distribution

Construction of the prior distribution for B the short dataset using the information coming from the long dataset A



Methodology

- we assume π^B belongs to a parametric family $\mathcal{F} = \{\pi_\lambda,\ \lambda \in \Lambda\}$
- we want to pick λ^B such that π^B is «close» to $\pi^A(\cdot|X^A)$

Let $T : \mathcal{F} \to \Lambda$ be an operator such that for every $\lambda \in \Lambda$,

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

Choice of the *prior* distribution on B

- we select λ^B «proportional to » $T[\pi^A(\cdot|y^A)]$ in the sense that

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

  where $K : \Lambda \to \Lambda$ is a diagonal operator

$$K = \mathrm{Diag}(\mathrm{k}_1, \ldots, \mathrm{k}_\mathrm{j}, \ldots)$$

- the operator K can be interpreted as a similarity operator between A and B (the diagonal components of K are hyperparameters of the prior)
- we give k_j a vague hierarchical prior distribution centred around q,
- the prior on q is vague and centred around 1.

Methodology (cont.)

Example 1 : Method of Moments

We assume that the elements of $\mathcal{F}$ can be identified via their m first moments.

$$T[\pi_\lambda] = F(\mathbb{E}(\theta), \ldots, \mathbb{E}(\theta^m)) = \lambda$$

The expression of λ^B then becomes

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

Example 2 : Conjugacy

We assume that $\mathcal{F}$ is the family of priors conjugated for the model.

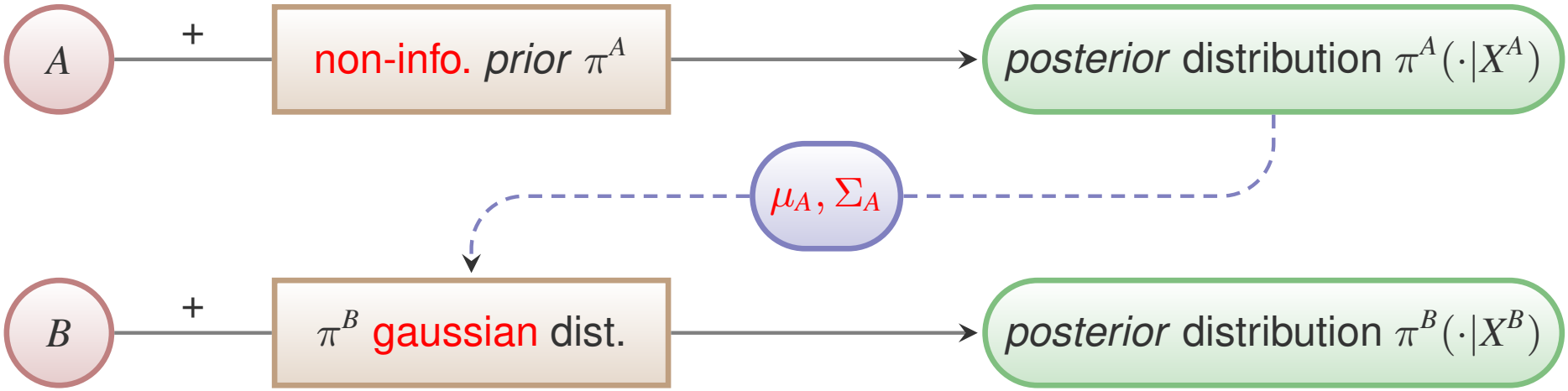
$$\pi^A \in \mathcal{F} \Rightarrow \pi^A(\cdot|X^A) \in \mathcal{F}$$

let $\lambda^A(X^A) \in \Lambda$ be the parameter of *posterior* distribution $\pi^A(\cdot|X^A)$
The expression of λ^B thus reduces to

$$\lambda^B = K\lambda^A(X^A)$$

Application to the model of electricity load

- Non informative prior for the the long dataset A
- $\pi^B \in \mathcal{F} = \{gaussian\ distributions\}$



Non-informative prior [long dataset A]

non-informative *prior* for A

- we consider *prior* distribution of the form : $\pi^A(\theta) = \pi(\eta)\pi(\sigma^2)$
  - $\star$ $\pi(\eta) \propto 1$
  - $\star$ $\pi(\sigma^2) \propto \sigma^{-2}$

Jeffreys prior for a standard regression model

posterior distribution

- the *posterior* distribution associated is well defined

$$\int_\Theta l(X_1, \ldots, X_N|\theta)\pi^A(\theta)\,\mathrm{d}\theta < +\infty$$

- Explicit forms of μ^A, Σ^A are not available $\rightsquigarrow$ MCMC algorithm

Informative *prior* distribution [short dataset B]

The hierarchical prior that we use is built as follows :

- the *prior* distribution is of the form :

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

- *prior* : the parameter of B are close to μ^A and Σ^A

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

$\star$ $\eta|k, l \sim \mathcal{N}(M^A k, l^{-1}\Sigma^A)$ $\qquad M^A = \text{Diag}\, \mu_A$

- hyperparameters k and l correspond to the similarity between A and B

$\star$ $k_j|q, r \sim \mathcal{N}(q, r^{-1}),\ l \sim \mathcal{G}(a_l, b_l)$ $\qquad a_l = b_l = 10^{-3}$

- hyperparameters q and r are more general indicators of how close A and B are,
  q corresponding to the mean of the coordinates of k
  r is the inverse-variance of the coordinates of k

$\star$ $q \sim \mathcal{N}(1, \sigma_q^2),\ r \sim \mathcal{G}(a_r, b_r)$ $\qquad \sigma_q^2 = 10^4,\ a_r = b_r = 10^{-6}$

Plan

Simulation

Dataset A.

- We simulated 4 years of daily data for A with parameters chosen to mimic the typical electricity load of France .
  - ⋆ 4 frequencies used for the truncated Fourier series
  - ⋆ 7 daytypes : one daytype for each day of the week
  - ⋆ 2 offsets to simulate the daylight saving time effect
- The temperatures (T_t) are those measured from September 1996 to August 2000 at 10 :00AM.

Dataset B. We simulated 1 year of daily data for B with parameters :

$$\begin{aligned} \text{seasonal} &: \alpha_i^B = k\alpha_i^A, \quad \alpha = (a, b, \omega) && \forall i = 1, \ldots, d_\alpha \\ \text{shape} &: \kappa_j^B = \kappa_j^A, && \forall j = 1, \ldots, d_\beta \\ \text{heating} &: g^B = kg^A, \quad u^B = u^A. \end{aligned}$$

We compare the quality of the bayesian predictions associated to

- the hierarchical prior
- the non-informative prior.

We simulate an extra year of daily data B to evaluate the prediction

The comparison criterion

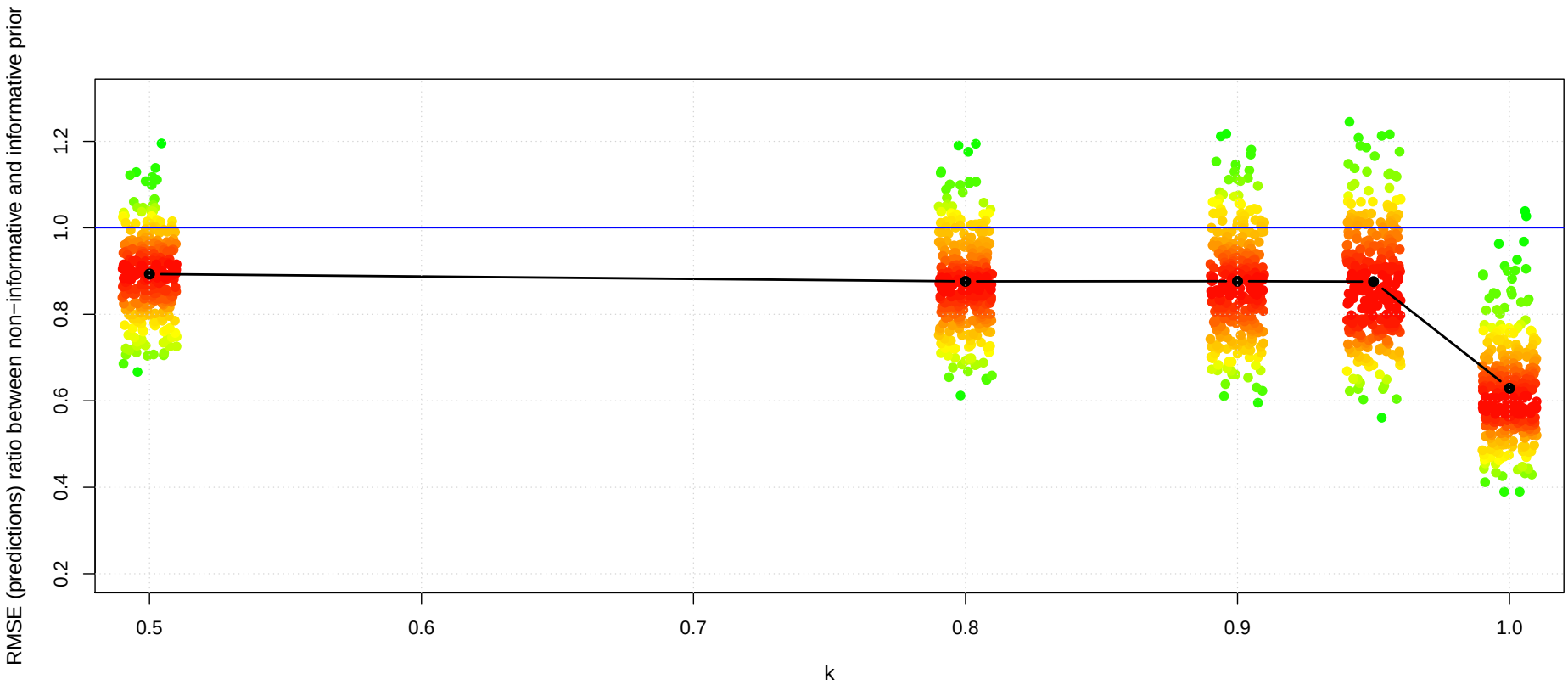
- let $X_t = f_t(\theta) + \epsilon_t$, the prediction error can be written as

$$X_t - \widehat{X}_t = \underbrace{[X_t - f_t(\theta)]}_{\text{noise } \epsilon_t} + \underbrace{[f_t(\theta) - \widehat{X}_t]}_{\text{difference w.r.t. the true model}}$$

- given that we want to validate our model on simulated data, we 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. :

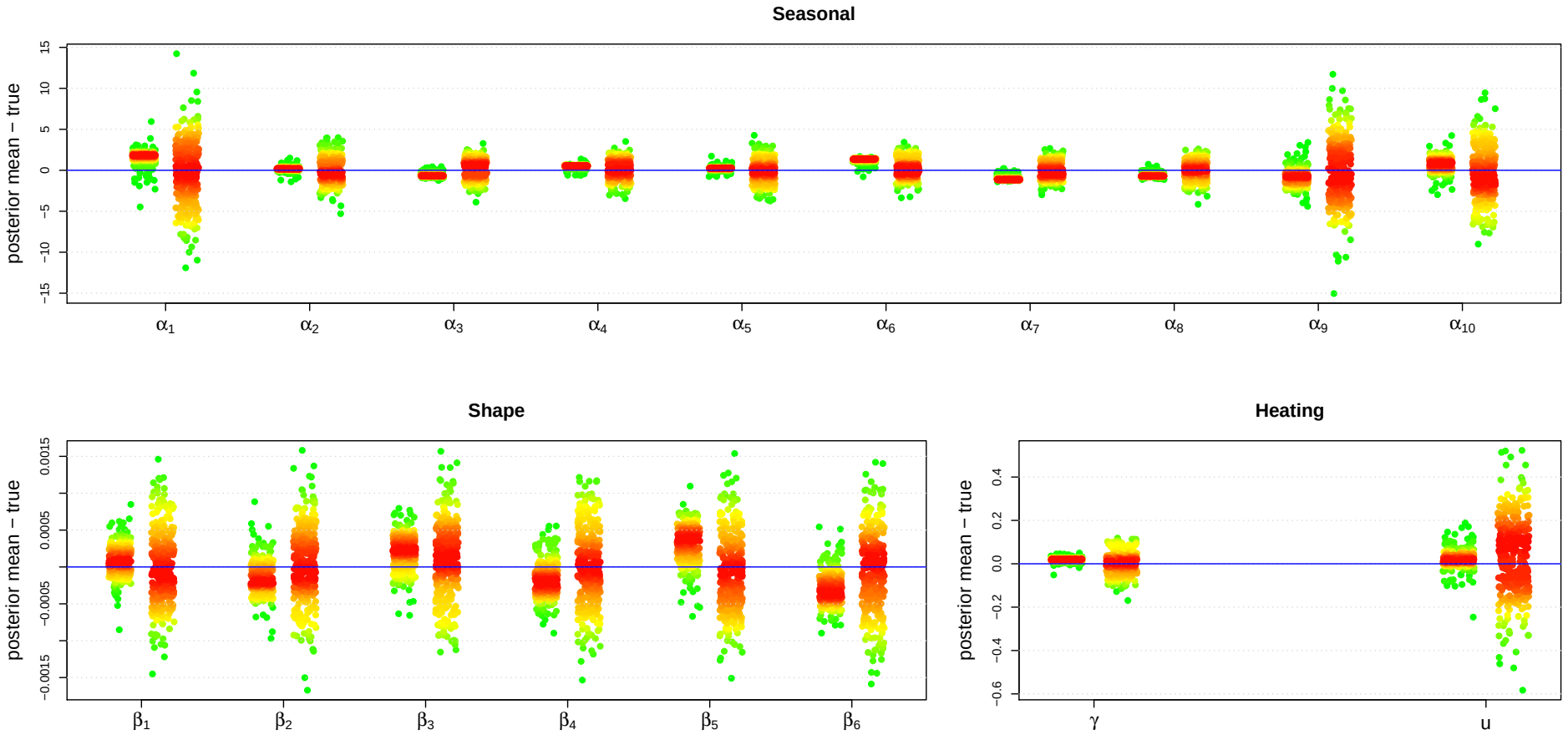
$$\sqrt{\frac{1}{365}\sum_{t=1}^{365}[f_t(\theta) - \widehat{X}_t]^2}.$$

Comparison : RMSE info. / RMSE non-info.



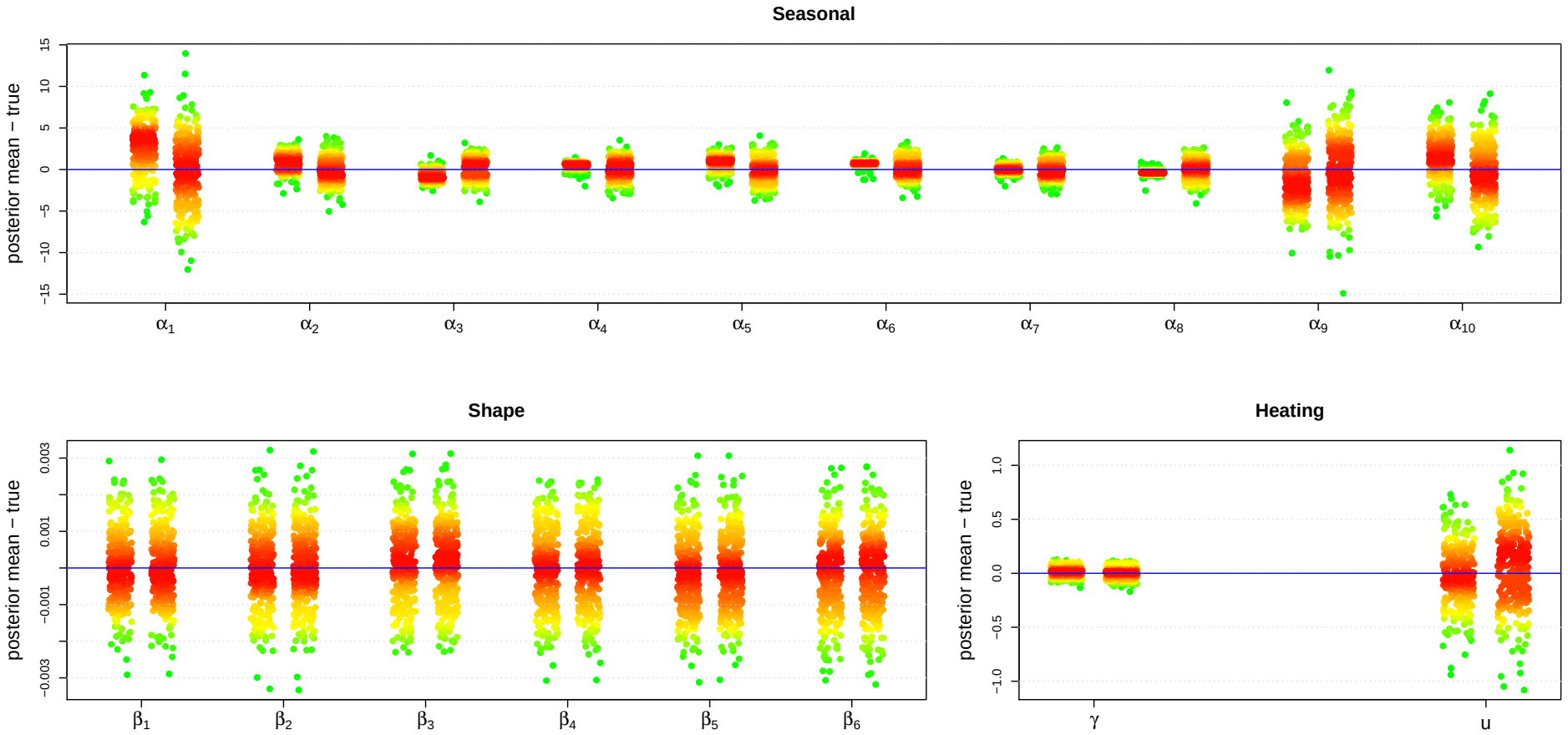
- abscissa : k, similarity coefficient between seasonality and heating gradient of populations A and B (other parameters remain untouched)
- ordinate : RMSE info. / RMSE non-info.
  - ⋆ whether similarity between A and B is ideal ($k = 1$) or not ($k \neq 1$), *prior* information improves predictions' quality

Parameters' estimation : info. vs non-info.



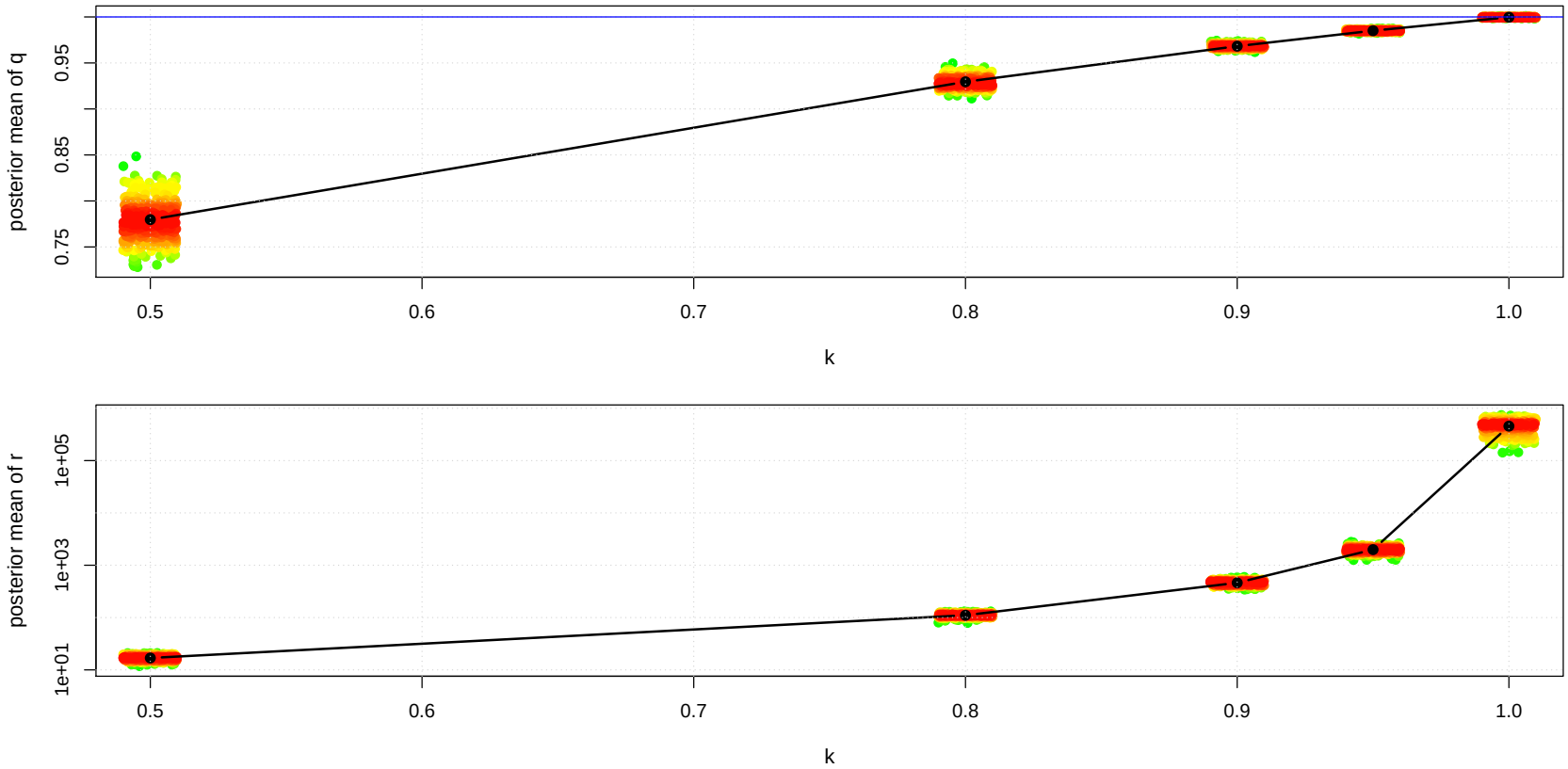
- 300 replications for the ideal situation $(k = 1)$
- ordinate : $\widehat{\theta} - \theta$ for the informative *prior* (left) and the non-informative *prior* (right)
  - ⋆ *prior* information improves the estimations a lot

Parameters' estimation : info. vs non-info. (cont.)



- 300 replications for a difficult situation $(k = 0.5)$
- ordinate : $\widehat{\theta} - \theta$ for the informative *prior* (left) and the non-informative *prior* (right)
  - ⋆ *prior* information improves the estimations only slightly

Hyperparameters' estimation (q and r) : $k_j \sim \mathcal{N}(q, r^{-1})$



- abscissa : k, similarity coefficient between seasonality and heating gradient of populations A and B (other parameters remain untouched)
- ordinate : Bayes estimator for q and r for the 5×300 replications considered
  - ⋆ q et r correspond to the mean and precision (inverse of variance) of the similarity coefficients k_j

Application

Dataset

- For A the dataset corresponds to a specific population in France frequently referred to as “non-metered”
- B_1 is a subpopulation of A
- B_2 covers the same people that A (ERDF)
- The sizes (in days) of the datasets are

⋆ population A : non metered (833 days)

⋆ population B_1 : res1+2 (177 +30 days)

⋆ population B_2 : non metered ERDF (121 +30 days)

Dataset

⋆ 4 frequencies used for the truncated Fourier series

⋆ 5 daytypes

⋆ 2 offsets

We kept the last 30 days of each B out of the estimation datasets and assessed the model quality over the predictions for those 30 days.

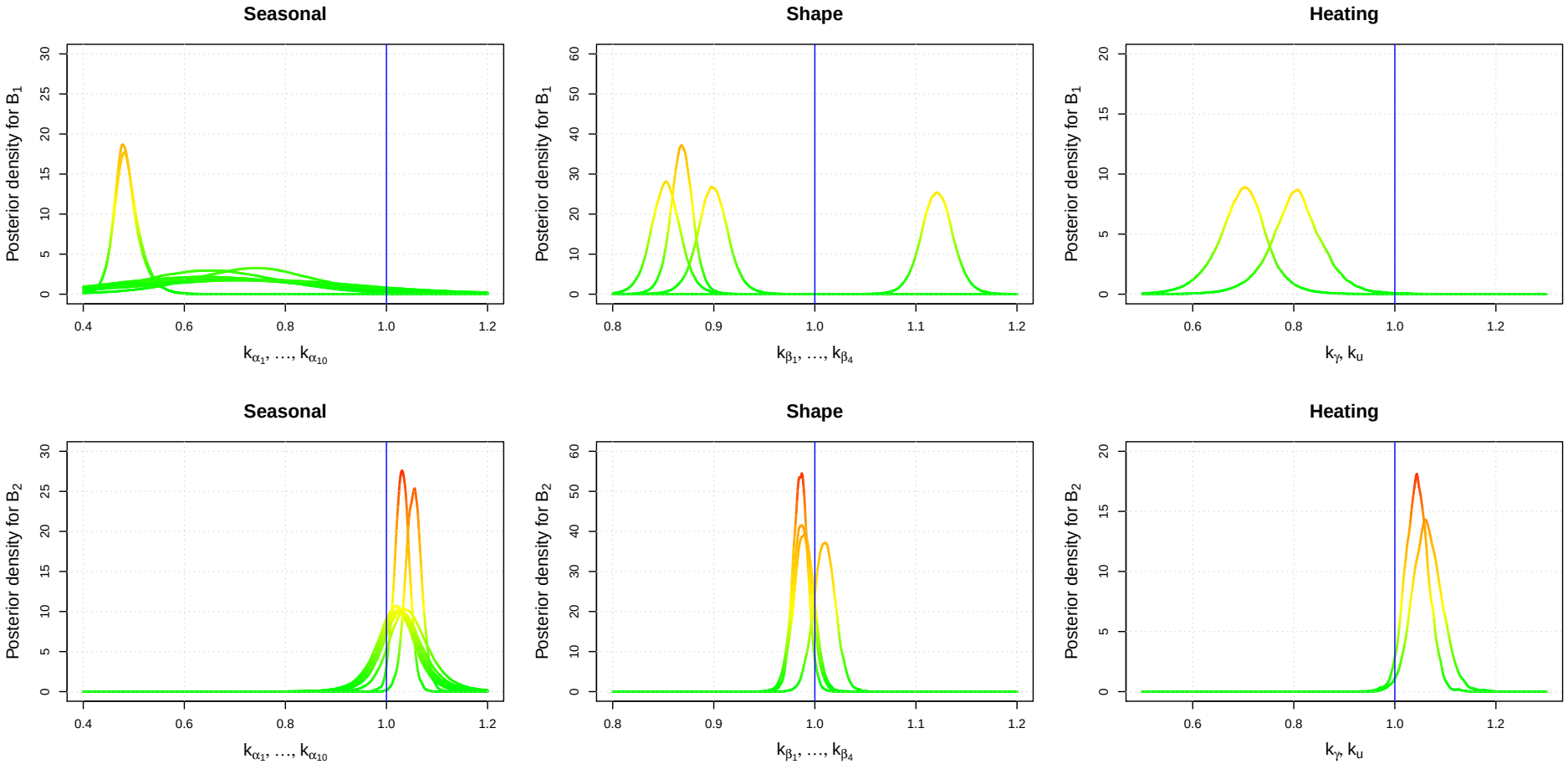
Results

B_1	non-informative	hierarchical	comparison
RMSE est.	775.93	786.97	$+1.42\%$
RMSE pred.	1863.25	894.00	-52.01%
MAPE est.	4.00	3.93	-0.07
MAPE pred.	19.37	9.30	-10.07

B_2	non-informative	hierarchical	comparison
RMSE est.	1127.60	1202.32	$+6.62\%$
RMSE pred.	2286.42	1339.14	-41.83%
MAPE est.	2.82	2.98	$+0.15$
MAPE pred.	8.65	3.48	-5.17

TABLE: Results for the dataset B_1 (top) and B_2 (bottom). RMSE is the "root mean square error" and MAPE is the "mean absolute percentage error". Both of these common measures of accuracy were computed on the estimation (est.) and prediction (pred.) parts of the two datasets.

Estimated posterior marginal distributions of k_i for the hierarchical method and for both datasets B_1 (upper row) and B_2 (lower row).



Estimation of hyperparameters $(\widehat{q}, \widehat{r}) = \begin{cases} (0.73, 24.48) & \text{on } B_1 \\ (1.02, 795.16) & \text{on } B_2 \end{cases}$

Plan

References I


T. Launay, A. Philippe and S. Lamarche (2011) Construction of an informative hierarchical prior distribution. Application to electricity load forecasting.
http://arxiv.org/abs/1109.4533v2


T. Launay, A. Philippe and S. Lamarche (2012) Consistency of the posterior distribution and MLE for piecewise linear regression
http://arxiv.org/abs/1203.4753