Bayesian methods for electricity load forecasting

Tristan Launay

To cite this version:

Tristan Launay. Bayesian methods for electricity load forecasting. Applications [stat.AP]. Université de Nantes, 2012. English. <tel-00766237>

HAL Id: tel-00766237

https://tel.archives-ouvertes.fr/tel-00766237

Submitted on 17 Dec 2012

UNIVERSITÉ DE NANTES
FACULTÉ DES SCIENCES ET TECHNIQUES

ÉCOLE DOCTORALE SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATHÉMATIQUES

Année : 2012 N° B.U. : | | | | | | | | | | |

Méthodes bayésiennes pour la prévision de consommation d'électricité

Thèse de Doctorat de l'Université de Nantes

Spécialité : MATHÉMATIQUES ET APPLICATIONS

Présentée et soutenue publiquement par

Tristan LAUNAY

Le 12 décembre 2012, devant le jury ci-dessous

Président du jury	:	Georges OPPENHEIM	(Université de Paris 11)
Rapporteurs	:	Siem Jan KOOPMAN	(Vrije Universiteit Amsterdam)
		Jean-Michel MARIN	(Université de Montpellier 2)
Examinateurs	:	Philippe CARMONA	(Université de Nantes)
		Sophie LAMARCHE	(EDF R&D)
		Frédéric LAVANCIER	(Université de Nantes)
		Nadia OUDJANE	(EDF R&D)
		Éric PARENT	(UMR AgroParisTech/INRA 518)
Directeur de thèse	:	Anne PHILIPPE	(Université de Nantes)
Laboratoire	:	Laboratoire Jean Leray	(UMR 6629 UN-CNRS-ECN)

N° E.D. :

Remerciements

> "And now it begins," said Ser Arthur Dayne, the Sword of the Morning. He unsheathed Dawn and held it with both hands. The blade was pale as milkglass, alive with light. "No," Ned said with sadness in his voice. "Now it ends."
>
> **George R. R. Martin**, *A Game of Thrones*

Qu'il me soit permis dans les quelques lignes à venir de remercier les nombreuses personnes qui ont contribué à l'élaboration des résultats présentés ici, de près ou de loin, de manière directe ou indirecte.

Je tiens à remercier en premier lieu les personnes qui m'ont encadré tout au long de ses trois dernières années. Merci Anne pour ta disponibilité et tes conseils avisés tant sur le fond que sur la forme : la lisibilité de ce manuscrit est en grande partie le fruit de tes interventions. Merci Sophie pour ta patience et ta bienveillance : l'amplitude de mon signal haute fréquence a diminué de façon régulière et notable grâce à toi. Merci à toutes deux pour votre implication sincère et continue tout au long de mes travaux.

Je souhaite également remercier Siem Jan Koopman et Jean-Michel Marin d'avoir accepté de rapporter cette thèse, ainsi que Philippe Carmona, Frédéric Lavancier, Georges Oppenheim, Nadia Oudjane et Eric Parent pour leur participation au jury.

Je remercie toutes les personnes du groupe R39 à EDF R&D que j'ai eu l'occasion de côtoyer, toujours patientes et prêtes à répondre à mes questions. Je n'oublierai pas nos nombreux échanges autour d'un repas ou d'une boisson chaude. Merci en particulier à Geoffray qui a guidé mes pas du côté obscur de la Force et sans qui certains calculs n'auraient certainement pas encore aboutis à l'heure actuelle. Merci spécial à Jairo, avec qui j'ai eu le plaisir de partager le bureau B305 : nos conversations m'ont aidé à prendre du recul sur beaucoup de choses. Hasta luego, amigo !

Je tiens aussi à remercier les personnes extérieures au groupe dont la présence et l'efficacité ont facilité mon parcours au sein de l'entreprise : notamment Lise et Claudine qui ont guidé mes pas dans le labyrinthe administratif et Véronique et Karim qui ont sauvé manuscrit, données, codes et résultats de l'abîme numérique (avec le sourire en plus).

Merci à l'ensemble des doctorants du Laboratoire de Mathématiques Jean Leray pour leur accueil toujours chaleureux lors de mes passages sporadiques à Nantes. Merci en particulier à Alexandre, Alexandre (l'autre !), Céline, Thomas, Vincent et Vivien pour leur bonne humeur, et les nombreuses discussions autour d'un jeu de cartes ou d'une partie d'échecs.

Merci à Paul et à Marie (il est loin le temps du bureau B314) pour les bons moments passés ensemble autour d'un verre.

Merci à ma mère et à mes frères pour leur présence toujours renouvelée, contre vents et marées, et leur soutien inconditionnel. Merci à Kina, tout court.

Table des matières

1 Introduction

Le présent manuscrit s'intéresse à la prévision de la consommation d'électricité par des méthodes de statistiques bayésiennes. Trois thèmes y sont développés, dans trois chapitres successifs, concernant le comportement asymptotique des estimateurs de Bayes (afin de valider l'inférence liée au modèle de part chauffage), la construction d'une loi a priori hiérarchique (pour améliorer les prévisions en situation d'historique court) et l'exploitation d'un modèle dynamique (dans le but de disposer de prévisions court terme en ligne). Ce premier chapitre propose donc une vue d'ensemble des différents concepts et objets autour de ces thèmes. Nous présentons en section 1.1 l'activité de prévision de consommation d'électricité, puis introduisons en section 1.2 les concepts clefs situés au coeur des méthodes bayésiennes dans un cadre plus général. Nous terminons ce chapitre en proposant un aperçu général des problématiques abordées et des résultats obtenus en section 1.3.

1.1 La prévision de consommation d'électricité

Dans cette section nous présentons rapidement les enjeux liés à la prévision de consommation. Nous décrivons également les caractéristiques majeures du signal de consommation d'électricité en France ainsi que le modèle principalement utilisé en opérationnel.

1.1.1 Les enjeux

La modélisation et la prévision de la consommation d'électricité à différents horizons (court terme moyen terme, ou long terme) représentent une activité clef pour le groupe EDF : en effet, afin d'éviter les risques physiques (black-out partiel ou total), ou financiers (pénalités financières) l'équilibre doit être maintenu constamment entre la demande et l'offre d'énergie sur le réseau électrique. L'optimisation des coûts de production étant un enjeu essentiel pour le groupe EDF, l'activité de prévision court terme est donc directement liée à la gestion des moyens de production qui sont aussi nombreux que variés et regroupent notamment des centrales nucléaires, des centrales thermiques, des parcs éoliens, etc. Les prévisions moyen terme et long terme permettent, entre autres choses, d'établir les plannings de maintenance des différents moyens de production, et sont également utilisées pour décider des investissements futurs (par exemple en vue d'augmenter la capacité de production).

L'ouverture du marché de l'électricité à la concurrence en France, pour les entreprises comme pour les particuliers, a récemment conduit à des évolutions dans le domaine de la prévision d'électricité. Ainsi le périmètre des clients EDF, jusqu'alors fixe et égal au périmètre France, est désormais soumis à des variations dues aux arrivées et départs de clients, ce qui rend la prévision de la demande EDF plus difficile et explique l'intérêt renouvelé du groupe envers le développement de méthodes innovantes pour prévoir un signal dont la stationnarité se trouve remise en question (voir par exemple les thèses de Cugliari, 2011; Dordonnat, 2009; Goude, 2008).

1.1.2 La consommation d'électricité en France

Nous illustrons les principales caractéristiques de la consommation nationale d'électricité en France à travers les figures 1.1 et 1.2. Le signal de consommation présente de manière générale trois cycles : un cycle annuel, un cycle hebdomadaire et un cycle journalier. La figure 1.1 représente le comportement moyen typique de la consommation d'électricité sur une année : la demande est globalement plus élevée durant l'hiver que durant l'été (à cause du développement important du chauffage électrique en France). L'impact de l'activité économique du pays est aisément repérable sur le graphique avec la présence de ruptures visibles, aux périodes de congés en Août ou durant la période de Noël. Nous distinguons également la présence d'un cycle hebdomadaire, reflet d'une demande d'électricité moindre les jours de weekends comparativement aux jours de semaine ouvrés.

La figure 1.2 (gauche) représente la consommation délectricité au cours de deux semaines de l'année 2005. La variation en niveau moyen entre les deux semaines est le simple reflet de la position de ces semaines dans l'année, l'une appartenant à l'été l'autre à l'hiver. Nous distinguons néanmoins la présence d'un cycle journalier, au sein du cycle hebdomadaire, dans les deux cas : de manière générale, la demande d'électricité est plus forte pendant la journée que pendant la nuit. Le cycle journalier semble toutefois différent entre l'hiver et l'été, avec la présence, en hiver, d'un pic de consommation plus marqué en soirée aux alentours de 19h00. L'effet de l'instant de la journée visible sur cette figure, conduit de manière générale à modéliser les 48 instants de façon indépendante.

Enfin, la figure 1.2 (droite) illustre le lien non linéaire qui existe entre la consommation d'électricité et la température extérieure. En deça d'une certaine température (appelée température seuil de chauffage, ou plus simplement seuil de chauffage) la consommation semble d'autant plus importante que la température est froide. Notons qu'un phénomène similaire se produit pour les températures chaudes, et correspond à un effet climatisation (par opposition à l'effet chauffage que nous venons de décrire). Nous constatons toutefois que l'effet climatisation est beaucoup moins marqué que l'effet chauffage à l'échelle nationale, car l'utilisation de climatiseurs est globalement moins répandue en France que celle du chauffage électrique.

1.1.3 Les modèles de prévision

Les approches utilisées dans la littérature pour modéliser et prévoir la consommation d'électricité sont très variées. Nous en mentionnons quelques-unes dans les lignes ci-dessous, avant de nous intéresser plus particulièrement au modèle principalement utilisé par le groupe EDF.

Etat de l'art

De nombreux travaux considèrent des séries temporelles univariées : Taylor (2003) construit un modèle basé sur un lissage double exponentiel pour la consommation d'électricité au Royaume-Uni et Taylor et al. (2006); Taylor and McSharry (2007) présentent une étude comparative des méthodes univariées pour différents jeux de données. D'autres travaux prennent en compte des variables exogènes Harvey and Koopman (1993) incluent la température extérieure dans leur modèle, à l'origine du modèle bayésien semi-paramétrique développé dans Smith (2000).

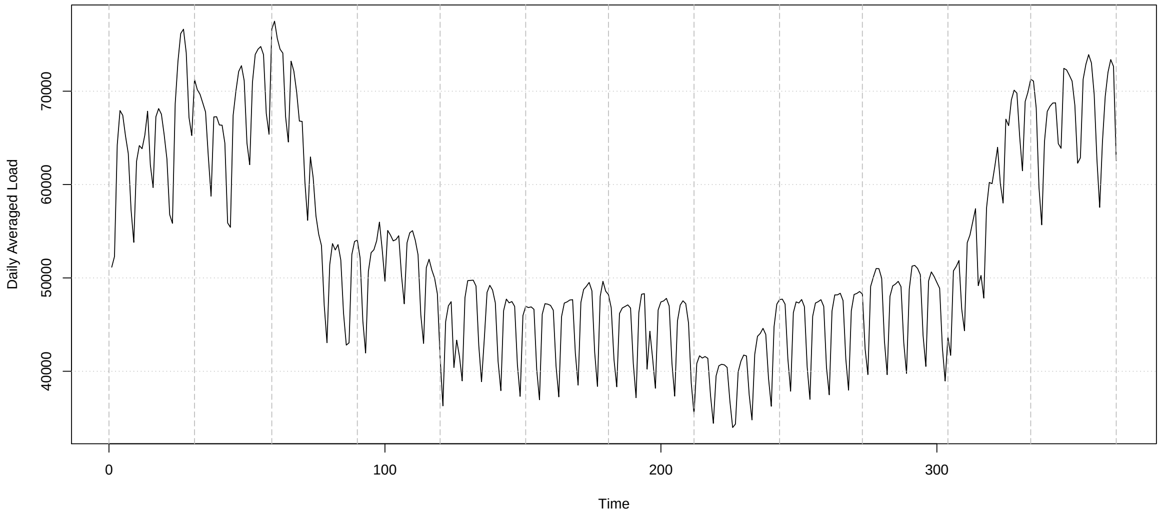


Figure 1.1: Consommation d'électricité moyenne journalière pour l'année 2005. Les lignes verticales en tirets marquent la séparation entre les différents mois.

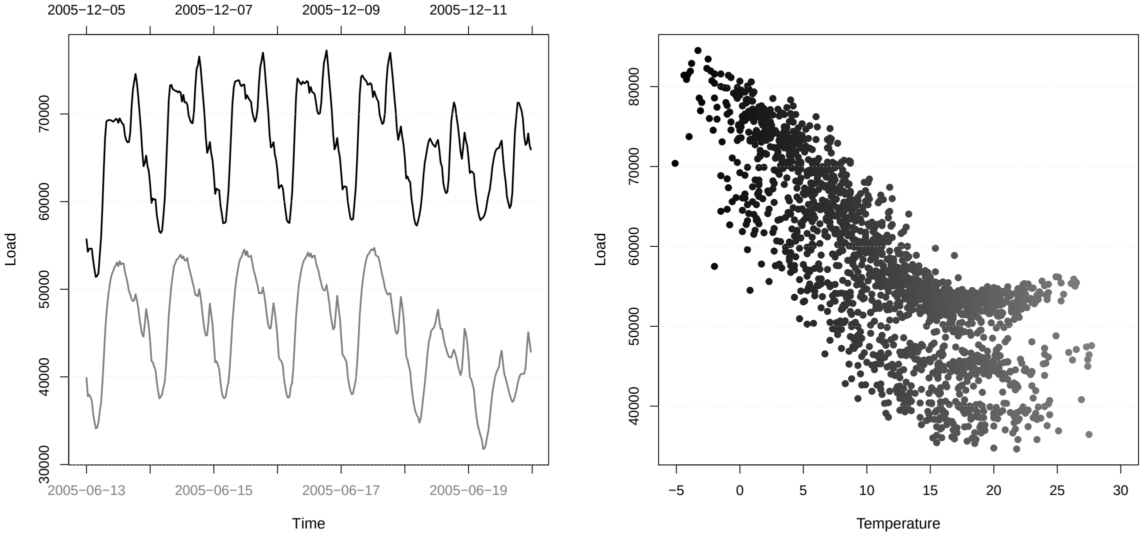


Figure 1.2: Consommation d'électricité instantanée sur deux semaines en 2005 (à gauche) et à 10h00 du matin en fonction de la température relevée sur la période 2002–2007 (à droite).

Des solutions alternatives à la modélisation univariées sont également proposées, telles que des modèles multi-équations : les différents instants de la journée partagent la même équation de modélisation mais les paramètres du modèles sont différents d'un instant à l'autre. Soares and Medeiros (2008) construisent ainsi un modèle horaire saisonnier auto-régressif pour leurs données et Ramanathan et al. (1997) construisent également un modèle indépendant pour chaque heure de la journée mais prenant en compte un effet de la température extérieure.

Les variables exogènes les plus fréquemment utilisées pour prévoir la consommation d'électricité sont des variables météorologiques : l'inclusion de variables météorologiques dans un modèle conduit à de nouvelles difficultés puisqu'il devient nécessaire de disposer de prévisions météorologiques fiables pour obtenir des prévisions de consommation d'électricité de bonne qualité. Pour les climats tempérés, le facteur météorologique le plus important est la température extérieure (voir par exemple Taylor and Buizza, 2003). Plus spécifiquement, pour la consommation d'électricité en France, l'importance de la température et de la nébulosité (couverture nuageuse) est soulignée dans Bruhns et al. (2005); Menage et al. (1988). Des modèles exploitant d'autres variables météorologiques sont également présentés dans Engle et al. (1986).

Cottet and Smith (2003); Smith and Kohn (2002) proposent d'utiliser une approche bayésienne pour l'estimation et la prévision de modèles de consommation d'électricité. Ainsi Smith and Kohn (2002) s'intéressent à l'estimation de matrices de covariances lacunaires à partir d'une loi a priori hiérarchique et utilisent sur une période de quinze jours ouvrés une modélisation indépendante des différents instants de la journée à laquelle ils rajoutent une structure d'auto-corrélation sur les erreurs (entre deux instants de la journée). Cottet and Smith (2003) développent pour leur part un modèle similaire en incluant des termes d'auto-régression sur les erreurs, dont ils restreignent les coefficients au domaine de stationnarité grâce au choix de la loi a priori et appliquent leur méthode sur un historique de trois ans de données.

Signalons enfin quelques travaux innovants récents au sein du groupe EDF en lien avec la thématique de prévision de consommation d'électricité. Goude (2008) présente une étude détaillée des techniques de mélange de prédicteurs, utilisées pour agrégér différentes prévisions en fonction de leurs performances les plus récentes. Dordonnat (2009) propose trois formes de modèles à espace d'états pour lesquels les paramètres du modèle sont autorisés à varier dans le temps, et Pierrot and Goude (2011) considèrent des modèles additifs généralisés. Cugliari (2011) opte quant à lui pour une approche non paramétrique reposant sur la transformation en ondelettes pour la prévision de la courbe de consommation journalière qui est considérée comme un processus hilbertien auto-régréssif à valeurs fonctionnelles.

Modélisation Eventail

Le modèle utilisé pour les prévisions court terme en opérationnel par le groupe EDF est basé sur le modèle présenté originellement dans Bruhns et al. (2005). Ce modèle (appelé Eventail) étant à l'origine des modèles considérés dans les différents chapitres de ce manuscrit, nous en décrivons les caractéristiques principales ci-dessous.

Le modèle Eventail est un modèle de régression non linéaire qui décrit au pas demi-horaire la consommation d'électricité $y_{t,i}$ du jour t à l'instant i. Les demi-heures (ou instants) sont supposées indépendantes et chacune est modélisée pour $t = 1, \dots, N$ et $i = 0, \dots, 47$, par

$$\begin{aligned}
y_{t,i} &= x^{\text{season}}_{t,i} + x^{\text{heat}}_{t,i} + x^{\text{cool}}_{t,i} + \epsilon_{t,i} \qquad (1.1)\\
x^{\text{season}}_{t,i} &= x^{\text{Fourier}}_{t,i} x^{\text{shape}}_{t,i}\\
x^{\text{Fourier}}_{t,i} &= \sum_{j=1}^{d_{11,i}} \left[z^{\cos}_{j,i} \cos\left(\frac{2j\pi}{365.25} t\right) + z^{\sin}_{j,i} \sin\left(\frac{2j\pi}{365.25} t\right)\right] + \sum_{j=1}^{d_{12,i}} \omega_{j,i} \mathbb{1}_{\Omega_{j,i}}(t),\\
x^{\text{shape}}_{t,i} &= \sum_{j=1}^{d_{2,i}} \psi_{j,i} \mathbb{1}_{\Psi_{j,i}}(t),\\
x^{\text{heat}}_{t,i} &= g^{\text{heat}}_i (T^{\text{heat}}_{t,i} - u^{\text{heat}}_i) \mathbb{1}_{[T^{\text{heat}}_{t,i}, +\infty[}(u^{\text{heat}}_i),\\
x^{\text{cool}}_{t,i} &= g^{\text{cool}}_i (T^{\text{cool}}_{t,i} - u^{\text{cool}}_i) \mathbb{1}_{]-\infty, T^{\text{cool}}_{t,i}]}(u^{\text{cool}}_i),
\end{aligned}$$

où les variables aléatoires $\epsilon_{1,i}, \dots, \epsilon_{N,i}$ sont supposées indépendantes et identiquement distribuées selon la loi normale $\mathcal{N}(0, \sigma_i^2)$ et où les températures $T^{\text{heat}}_{t,i}$ et $T^{\text{cool}}_{t,i}$ sont définies à partir de la température extérieure brute $T^{\text{raw}}_{t,i}$ et de lissages exponentiels comme

$$\begin{aligned}
T^{\text{heat}}_{n,i} &= \alpha^{\text{heat}}_i \cdot T^{\vartheta^{\text{heat}}_{1,i}}_{n,i} + \beta^{\text{heat}}_i \cdot T^{\vartheta^{\text{heat}}_{2,i}}_{n,i} + (1 - \alpha^{\text{heat}}_i - \beta^{\text{heat}}_i) \cdot T^{\text{raw}}_{n,i},\\
T^{\text{cool}}_{n,i} &= \alpha^{\text{cool}}_i \cdot T^{\vartheta^{\text{cool}}_i}_{n,i} + (1 - \alpha^{\text{cool}}_i) \cdot T^{\text{raw}}_{n,i}.
\end{aligned}$$

où les lissages exponentiels de la température brute sont définis, pour un paramètre ϑ donné, par

$$\begin{aligned}
T^{\vartheta}_{0,0} &= T^{\text{raw}}_{0,0}\\
T^{\vartheta}_{n,i} &= \vartheta \cdot T^{\vartheta}_{n,i-1} + (1 - \vartheta) \cdot T^{\text{raw}}_{n,i}
\end{aligned}$$

en utilisant la notation $T^{\vartheta}_{n,-1} = T^{\vartheta}_{n-1,47}$.

Les paramètres du modèle à estimer sont de manière générale (il arrive parfois que certains de ces paramètres soient fixés à des valeurs prédéfinies) pour $i = 0, \dots, 47$

$$\sigma_i, \underbrace{z^{\cos}_{j,i}, z^{\sin}_{j,i}, \omega_{j,i}, \psi_{j,i}}_{x^{\text{season}}}, \underbrace{g^{\text{heat}}_i, u^{\text{heat}}_i, \alpha^{\text{heat}}_i, \beta^{\text{heat}}_i, \vartheta^{\text{heat}}_{1,i}, \vartheta^{\text{heat}}_{2,i}}_{x^{\text{heat}}}, \underbrace{g^{\text{cool}}_i, u^{\text{cool}}_i, \alpha^{\text{cool}}_i, \vartheta^{\text{cool}}_i}_{x^{\text{cool}}}.$$

La composante x^{Fourier} vise à capturer le comportement saisonnier moyen de la consommation d'électricité (i.e. le motif annuel) à partir d'une série de Fourier tronquée (entre 4 et 6 fréquences sont en général retenues) et d'une somme de fonctions indicatrices destinés à représenter les ruptures dans le signal à partir de partitions $(\Omega_{j,i})_{j \in \{1,\dots,d_{12,i}\}}$ du calendrier. Ces partitions servent à spécifier les périodes de congés ou les changements d'heure légale.

Le rôle de la composante x^{shape} est de capturer les cycles hebdomadaire et journalier en permettant l'ajustement journalier du comportement moyen (modélisé par x^{Fourier}). Cette composante repose sur l'utilisation de formes de jours $\psi_{j,i}$ à estimer, qui dépendent de types de jours spécifiés par d'autres partitions $(\Psi_{j,i})_{j \in \{1,\dots,d_{2,i}\}}$ du calendrier. Ces partitions servent à distinguer les jours ouvrés, des jours de fin de semaine ou encore des jours fériés. Pour des raisons d'identifiabilité, les formes de jours respectent pour tout $i = 0, \dots, 47$ la contrainte $\sum_{j=1}^{d_{2,i}} \psi_{j,i} = 1$.

Enfin, les composantes x^{heat} et x^{cool} ont pour but de modéliser la relation non linéaire existant entre la consommation d'électricité et la température, principalement à partir de seuils u^{heat}_i et u^{cool}_i et de gradients g^{heat}_i et g^{cool}_i qui correspondent aux intensités respectives des effets chauffage et climatisation.

1.2 LES MÉTHODES BAYÉSIENNES

Cette section présente un résumé des idées et outils élémentaires nécessaires au développement et à la mise en oeuvre de l'inférence bayésienne.

1.2.1 Le paradigme bayésien

Nous présentons dans cette section, une introduction succincte à l'inférence bayésienne. Nous restreignons cette présentation au cas des modèles paramétriques (voir Robert, 2007, par exemple) qui représentent le cadre de travail de ce manuscrit et laissons de côté les modèles non paramétriques (voir Ghosh and Ramamoorthi, 2003). Nous cherchons donc à mener l'inférence sur un paramètre $\theta \in \Theta$ inconnu, avec $\Theta \subset \mathbb{R}^d$ à partir de n observations $x_1, \ldots, x_n$ provenant d'un modèle statistique i.e. générées à partir d'une densité de probabilité $f(x|\theta)$ supposée connue. Par la suite, nous ferons souvent référence à $f(x|\theta)$ en tant que fonction de θ sous le nom de vraisemblance.

Approche fréquentiste vs. approche bayésienne

Une approche répandue en statistique dite fréquentiste (par opposition à la statistique bayésienne) repose sur l'utilisation de la méthode du maximum de vraisemblance. La méthode du maximum de vraisemblance consiste à estimer le paramètre θ inconnu par la valeur de θ (si elle existe) qui maximise la vraisemblance des observations, i.e. étant donné $x = (x_1, \ldots, x_n)$ le vecteur des observations, l'estimateur du maximum de vraisemblance $\widehat{\theta}_{\mathrm{MV}}$ est donné par

$$\widehat{\theta}_{\mathrm{MV}} = \arg\max_{\theta \in \Theta} f(x|\theta). \tag{1.2}$$

La maximisation de la fonction de vraisemblance peut toutefois poser problème en pratique, particulièrement dans le cas d'un espace Θ de grande dimension ou fortement contraint. L'approche du maximum de vraisemblance reste néanmoins séduisante dans la mesure où l'estimateur (1.2) considéré bénéficie de propriétés asymptotiques recherchées (comme la consistance et l'efficacité, la normalité asymptotique, etc.) dans de nombreux cas (voir DasGupta, 2008).

En statistique bayésienne, l'incertitude sur le paramètre θ d'un modèle statistique est modélisée par une loi de probabilité π appelée loi a priori et la vraisemblance $f(x|\theta)$ est interprétée comme la loi des observations conditionnellement au paramètre. Le paramètre θ en statistique bayésienne n'est donc plus considéré inconnu, comme en statistique fréquentiste, mais incertain puisqu'il est assimilé à une variable aléatoire. Rappelons la règle de Bayes (1763), dans sa version variables aléatoires : soient X et Y deux variables aléatoires (telles que la loi du couple (X, Y) possède une densité) de densité conditionnelle $f(x|y)$ et marginale $g(y)$ alors la densité conditionnelle de Y sachant X est donnée par

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)\,\mathrm{d}y}. \tag{1.3}$$

Le résultat d'inversion (1.3) permet de mener l'inférence à partir de la loi du paramètre θ conditionnellement aux observations x, appelée loi a posteriori et définie par

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,\mathrm{d}\theta}. \tag{1.4}$$

Estimateur bayésien

Un estimateur $\delta^*(x)$ est un estimateur de Bayes sous le coût $L(\theta,\delta)$ s'il minimise le risque bayésien c'est-à-dire :

$$\delta^* = \arg\min_{\delta} \iint L\big(\theta,\delta(x)\big) f(x|\theta)\pi(\theta)\,\mathrm{d}x\,\mathrm{d}\theta.$$

Pour le coût L^2 défini par $L(\theta,\delta) = (\theta-\delta)^2$ (utilisé dans la suite de ce manuscrit), l'espérance de la loi posteriori définie en (1.4) est un estimateur de Bayes :

$$\widehat{\theta} = \int \theta\pi(\theta|x)\,\mathrm{d}\theta, \tag{1.5}$$

et il s'avère qu'il n'en existe pas d'autre (c.f. Robert, 2007, pour la démonstration). D'autres estimateurs sont envisageables selon les problèmes et la fonction de coût considérés : la médiane a posteriori est par exemple un estimateur de Bayes pour le coût L^1 défini par $L(\theta,\delta) = |\theta-\delta|$.

Régions de crédibilité

Intervalles de crédibilité. Pour un paramètre θ unidimensionnel, la définition d'un intervalle de crédibilté (terme utilisé pour marquer la distinction avec les intervalles de confiance obtenus en statistique fréquentiste), est défini à partir des quantiles de la loi a posteriori (1.4), le quantile $q^\pi_\alpha(x)$ d'ordre $\alpha\in]0,1[$ vérifiant bien entendu la relation

$$\int_{-\infty}^{q^\pi_\alpha(x)} \pi(\theta|x)\,\mathrm{d}\theta = \alpha. \tag{1.6}$$

Les intervalles de crédibilité les plus souvent considérés sont les intervalles bilatéraux symétriques de niveau $1-\alpha$ qui sont de la forme

$$\Big]q^\pi_{\frac{\alpha}{2}}(x),\, q^\pi_{1-\frac{\alpha}{2}}(x)\Big[.$$

Régions HPD. Il est également possible de déterminer d'autres régions de crédibilité, appelées régions HPD (highest posterior density) de plus haute densité a posteriori. Ces régions $Q^\pi_{1-\alpha}(x)$ de niveau $1-\alpha$ sont de la forme

$$Q^\pi_{1-\alpha}(x) = \left\{\theta;\, \pi(\theta|x) \geqslant k^\pi_{1-\alpha}(x)\right\},$$

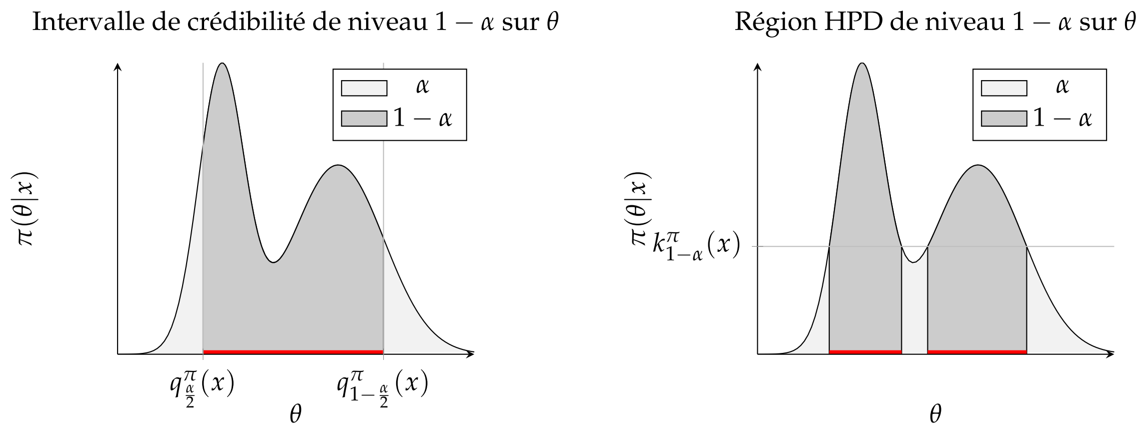
où $k^\pi_{1-\alpha}(x)$ vérifie la relation

$$\int \mathbb{1}_{\{\theta;\,\pi(\theta|x)\geqslant k^\pi_{1-\alpha}(x)\}}(\theta)\pi(\theta|x)\,\mathrm{d}\theta = 1-\alpha.$$

À l'inverse des intervalles de crédibilité, ces régions ne sont pas nécessairement connexes, comme le montre l'illustration fournie en Figure 1.3. Il est important de noter que la notion de région HPD s'étend naturellement au cas où le paramètre θ est multidimensionnel.

Probabilité de couverture fréquentiste d'une région de crédibilité. Une région de crédibilité a posteriori $R^\pi_{1-\alpha}(x)$ de niveau $1-\alpha$ sur θ, i.e. telle que

$$\int \mathbb{1}_{R^\pi_{1-\alpha}(x)}(\theta)\pi(\theta|x)\,\mathrm{d}\theta = 1-\alpha, \tag{1.7}$$


Intervalle de crédibilité de niveau $1-\alpha$ sur θ

Région HPD de niveau $1-\alpha$ sur θ

Figure 1.3: Exemples (en rouge) d'un intervalle de crédibilité bilatéral de niveau $1-\alpha$ sur θ (gauche) et d'une région HPD de niveau $1-\alpha$ sur θ (droite) construits à partir d'une même loi a posteriori $\pi(\theta|x)$.

définit également une région de confiance au sens fréquentiste, dont la probabilité de couverture fréquentiste (ou niveau fréquentiste) est donnée par

$$\mathbb{P}_{\theta}\left(R^{\pi}_{1-\alpha}(x) \ni \theta\right) = \int \mathbb{1}_{R^{\pi}_{1-\alpha}(x)}(\theta) f(x|\theta)\,\mathrm{d}x. \tag{1.8}$$

La probabilité de couverture fréquentiste diffère en général de $1-\alpha$ mais Datta and Mukerjee (2004) démontrent pour les intervalles de crédibilité unilatéraux et les régions HPD sous des hypothèses générales de régularité que

$$\mathbb{P}_{\theta}\left(R^{\pi}_{1-\alpha}(x) \ni \theta\right) = 1-\alpha+\mathrm{o}(1),$$

quand le nombre n d'observations tend vers $+\infty$.

1.2.2 La loi a priori

Le choix de la loi a priori représente une étape cruciale dans l'analyse statistique bayésienne, puisqu'elle influence directement le reste de l'inférence. L'information a priori, quand elle est disponible, n'est en général pas formulée en termes précis ou même statistiques : par exemple l'information a priori "θ appartient à l'intervalle $[-1, 1]$ avec 95% de chance" ne détermine en rien la forme de la loi a priori π qui peut être choisie indifféramment gaussienne, de Cauchy, uniforme, etc.

Lois conjuguées

Une famille $\mathcal{F}$ de lois de probabilité sur Θ est dite conjuguée pour la vraisemblance $f(x|\theta)$ si pour toute loi a priori $\pi \in \mathcal{F}$, la loi a posteriori $\pi(\cdot|x)$ appartient également à $\mathcal{F}$. Des exemples pour des modèles usuels sont présentés dans Robert (2007).

Les familles de lois conjuguées sont souvent considérées en premier lieu pour choisir la loi a priori car elles permettent d'effectuer les estimations sans avoir recours à des techniques complexes d'approximation numérique : si la famille conjuguée est une famille paramétrée $\mathcal{F} = \mathcal{F}_{\alpha \in A}$, la loi a posteriori appartenant à $\mathcal{F}_{\alpha \in A}$, son calcul se résume à une mise à jour du paramètre $\alpha \in A$. En particulier, l'accès aux quantités telles que l'espérance, la variance ou les intervalles de crédibilité a posteriori devient presque immédiat. Lorsqu'une information a priori est disponible, elle sert généralement à choisir le paramètre $\alpha \in A$ pour fixer le choix de la loi a priori au sein de la famille conjuguée. Cette facilité à mener l'estimation a toutefois un prix, puisque le choix de la loi a priori demeure restreint à la famille $\mathcal{F}_{\alpha \in A}$ considérée.

Lois hiérarchiques

D'une manière très générale, dès que le choix de la loi a priori est restreint à une famille paramétrée $\mathcal{F}_{\alpha \in A}$, que cette famille soit ou non conjuguée pour le modèle, il est nécessaire de choisir le paramètre $\alpha \in A$. Dans le cas où l'information a priori ne permet pas de fixer α, l'approche hiérarchique modélise le manque d'information sur α à l'aide d'une distribution a priori sur ce nouveau paramètre du modèle (alors appelé hyperparamètre). La loi a priori hiérarchique ainsi construit s'exprime alors souvent sous la forme :

$$\pi(\theta, \alpha) = \pi(\theta|\alpha)\pi(\alpha).$$

Le principe de l'approche hiérarchique peut aussi s'étendre à α lui-même dont la loi a priori peut dépendre d'un nouvel hyperparamètre, etc. L'utilisation de lois a priori hiérarchiques conduit à des estimateurs plus robustes, comme illustré dans Congdon (2003, 2010), au sens où l'inférence menée est moins sensible au choix des paramètres fixés par l'utilisateur.

Lois non informatives

En l'absence d'information a priori, le choix de la loi a priori s'effectue parmi les lois a priori dites non informatives puisqu'elles minimisent, en un certain sens, l'influence de la loi a priori sur la loi a posteriori. Nous ne présentons ici que quelques-unes des possibilités listées dans Kass and Wasserman (1996). Notons qu'il peut arriver que le choix de π considéré comme loi a priori ne définisse qu'une mesure positive et non une loi de probabilité sur Θ, i.e. $\pi(\theta) > 0$ pour tout $\theta \in \Theta$ et

$$\int \pi(\theta)\,\mathrm{d}\theta = +\infty.$$

Le cadre bayésien s'étend toutefois à de tels choix de lois a priori, dites impropres, dès lors que la loi (1.4) est bien définie, c'est-à-dire dès que

$$\int f(x|\theta)\pi(\theta)\,\mathrm{d}\theta < +\infty.$$

Loi de Laplace. La loi a priori de Laplace (1774) correspond au choix

$$\pi(\theta) \propto \mathbb{1}_{\Theta}(\theta).$$

En fonction de l'ensemble Θ des paramètres, nous retrouvons alternativement une loi uniforme Θ ou une loi impropre. Le choix de la loi de Laplace peut sembler naturel car aucune valeur de paramètre n'est a priori favorisée par rapport à une autre mais cette loi n'est pas invariante par reparamétrisation. En effet, si la reparamétrisation $\eta = g(\theta)$ est considérée (g étant une bijection) et si la loi de Laplace est choisie comme loi a priori sur θ, alors par changement de variable

$$\pi(\theta) \propto 1 \Longrightarrow \widetilde{\pi}(\eta) \propto \left| \frac{\mathrm{d}}{\mathrm{d}\eta} g^{-1}(\eta) \right|,$$

où $\tilde{\pi}$ désigne la loi a priori sur $\eta = g(\theta)$ correspondante. Bien qu'aucune information a priori ne soit disponible sur η, puisqu'aucune information a priori n'est disponible sur θ, le choix de loi a priori sur η n'est donc plus (en général) la loi de Laplace, et le choix de la loi a priori semble donc dépendre de la formulation même du problème.

Loi de Jeffreys. Jeffreys (1946) propose une loi a priori qui répond à la demande d'invariance par reparamétrisation. L'approche repose sur l'information de Fisher du modèle supposé régulier (voir Lehmann and Casella, 1998) définie pour $\theta \in \Theta \in \mathbb{R}^d$ comme la matrice $I(\theta)$ dont les coefficients sont donnés pour $1 \leqslant i, j \leqslant d$ par

$$I_{ij}(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta_i} \right) \left(\frac{\partial \log f(X|\theta)}{\partial \theta_j} \right) \right] \tag{1.9}$$

La loi non informative de Jeffreys est définie par

$$\pi^*(\theta) \propto \det^{\frac{1}{2}} I(\theta).$$

Clarke and Barron (1990) démontrent, dans le cas d'observations indépendantes et identiquement distribuées (i.i.d.), que ce choix minimise l'influence de la loi a priori sur la loi a posteriori au sens où elle maximise la divergence de Kullback-Leibler entre ces deux lois (voir également Philippe and Rousseau, 2002, pour une extension de ce résultat au cadre de la longue mémoire).

Lois de référence. Bernardo (1979) propose de construire des lois dite de référence : dans cette approche, les coordonnées sont regroupées par blocs sur lesquels un ordre est fixé et la loi a priori de référence est construite de manière conditionnelle. Le choix du nombre de blocs de coordonnées ainsi que leurs compositions et l'ordre qui leur est associé influencent donc la construction de la loi a priori de référence. Par exemple, pour $\theta = (\theta_1, \theta_2)$, où θ_1 désigne le paramètre d'intérêt et θ_2 le paramètre de nuisance, alors la loi a priori de référence est calculée en définissant d'abord $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$ conditionnellement à θ_1, puis $\pi(\theta_1)$ comme la loi Jeffreys associée à

$$\tilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\pi(\theta_2|\theta_1)\,\mathrm{d}\theta_2.$$

Les lois de référence constituent une généralisation de la loi de Jeffreys au sens où elles demeurent invariantes par reparamétrisation au sein de chaque bloc de coordonnées.

Lois "matching". La dernière approche que nous mentionnons ici consiste à choisir une loi a priori de manière à ce que la couverture bayésienne de certaines régions de crédibilité (1.7) (intervalles de crédibilité, ou bien régions HPD) coïncide (d'où le nom "matching") avec la couverture fréquentiste associée (1.8) jusqu'à un certain degré d'approximation (voir Datta and Mukerjee, 2004, pour un rassemblement des principaux résultats). Notant θ_1 la première coordonnée du vecteur $\theta \in \Theta$ avec $\Theta \subset \mathbb{R}^d$, et $q_\alpha^\pi(x)$ le quantile d'ordre α de la loi a posteriori marginale $\pi(\theta_1|x)$, Datta and Mukerjee (2004) démontrent par exemple sous des hypothèses générales de régularité l'approximation

$$\mathbb{P}_\theta\big(\theta_1 \leqslant q_\alpha^\pi(x)\big) = \alpha + \mathrm{O}\left(n^{-\frac{1}{2}}\right),$$

pour tout $\alpha \in]0,1[$. Une condition nécessaire et suffisante pour caractériser les lois a priori "matching" pour les quantiles a posteriori de θ_1 est également établie sous la forme de l'équation aux dérivées partielles

$$\sum_{j=1}^{d} \frac{\partial}{\partial \theta_j} \left\{ \pi(\theta) I^{j1}(\theta) \left(I^{11}(\theta) \right)^{-\frac{1}{2}} \right\} = 0, \tag{1.10}$$

où $I^{kl}(\theta)$ désigne le coefficient de la ligne k et de la colonne l de l'inverse de la matrice d'information de Fisher $I(\theta)$. Dans le cas d'un paramètre θ unidimensionnel (1.10) se réduit à l'équation différentielle

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \left\{ \pi(\theta) I(\theta)^{-\frac{1}{2}} \right\} = 0. \tag{1.11}$$

dont l'unique solution est la loi a priori de Jeffreys $\pi^*(\theta)$. Cette dernière vérifie l'approximation

$$\mathbb{P}_\theta \left(\theta \leqslant q_\alpha^{\pi^*}(x) \right) = \alpha + \mathrm{O}\left(n^{-1} \right),$$

pour tout $\alpha \in]0,1[$, améliorant ainsi (pour les intervalles de confiance unilatéraux) la vitesse de convergence de la couverture fréquentiste vers la couverture bayésienne d'un facteur $n^{-\frac{1}{2}}$.

1.2.3 Approximation numérique

Comme nous l'avons vu précémment, l'utilisation des lois a priori conjuguées permet le calcul explicite des quantités d'intérêt telles que l'espérance a posteriori. Ce calcul n'est en règle générale pas possible pour un autre choix de loi a priori, même pour des modèles simples. Nous présentons ci-dessous les différents problèmes d'approximation qui se posent dans le cas général ainsi que diverses solutions mises en oeuvre en pratique.

Les objets à estimer

Notons φ la loi d'intérêt. Dans le cadre bayésien, φ représente le plus souvent la loi a posteriori $\varphi(\theta) = \pi(\theta|x)$ définie en (1.4) mais peut également représenter la loi prédictive, définie par

$$p(x'|x) = \int f(x'|\theta)\pi(\theta|x)\,\mathrm{d}\theta. \tag{1.12}$$

Trois problèmes se présentent de manière générale : l'estimation de loi φ, l'estimation des quantiles de φ et le calcul d'intégrales de la forme

$$I = \mathbb{E}^\varphi[h(\theta)] = \int h(\theta)\varphi(\theta)\,\mathrm{d}\theta, \tag{1.13}$$

avec $h \in \mathrm{L}^1(\varphi)$.

L'approximation numérique de ces trois quantités par méthodes de Monte Carlo repose sur la simulation de variables aléatoires distribuée suivant la loi d'intérêt φ ou une loi approchée. Nous détaillons ci-après quelques-unes des méthodes stochastiques, largement utilisées dans la littérature, qui traitent de ce problème.

Simulation exacte

Lorsque la simulation exacte de variables aléatoires distribuées suivant la loi φ est disponible, la fonction de répartition empirique et les quantiles empiriques associés convergent vers la fonction de répartition et les quantiles associés à la loi φ (voir van der Vaart, 2000, par exemple).

Pour le calcul d'intégrales du type (1.13), il est possible d'utiliser la méthode de Monte Carlo que nous présentons ci-dessous avant d'aborder les techniques usuelles de simulation exacte.

Intégration de Monte Carlo. La Loi Forte des Grands Nombres (LFGN) assure que si $\theta^1, \dots, \theta^M$ sont des variables i.i.d. distribuées suivant la loi de densité $\varphi(\theta)$ alors

$$\frac{1}{M}\sum_{j=1}^{M} h(\theta^j) \xrightarrow{p.s.} \mathbb{E}^{\varphi}[h(\theta)]. \tag{1.14}$$

Pour $h \in \mathrm{L}^1(\varphi)$,

$$\widehat{I}_M^{\mathrm{MC}} = \frac{1}{M}\sum_{j=1}^{M} h(\theta^j) \tag{1.15}$$

définit donc un estimateur de (1.13), appelé estimateur de Monte Carlo, presque sûrement consistant, trivialement sans biais. Pour $h \in \mathrm{L}^2(\varphi)$, cet estimateur vérifie un théorème centrale limite (TCL)

$$\sqrt{M} \cdot \frac{\widehat{I}_M^{\mathrm{MC}} - I}{\widehat{\sigma}_M} \xrightarrow{d} N(0,1), \tag{1.16}$$

où $\widehat{\sigma}_M^2$ est l'estimateur de Monte Carlo de la variance asymptotique i.e.

$$\widehat{\sigma}_M^2 = \frac{1}{M}\sum_{j=1}^{M} h^2(\theta^j) - \left\{\frac{1}{M}\sum_{j=1}^{M} h(\theta^j)\right\}^2. \tag{1.17}$$

Techniques de simulation. Bien que le problème de la simulation d'une variable aléatoire suivant une loi d'intérêt φ fixée soit un problème très général (voir Devroye, 1986, par exemple), nous choisissons de ne présenter que la simulation directe et la simulation par acceptation-rejet (même si les solutions envisageables ne se limitent pas à ces seuls choix) : en effet, à l'exception de quelques cas très particuliers (par exemple en situation de conjugaison), la simulation exacte ne représente pas une solution viable pour simuler des variables aléatoires suivant la densité a posteriori $\varphi(\theta) = \pi(\theta|x)$.

Algorithme 1.1 (Méthode d'inversion). Pour simuler θ de fonction de répartition F

1. Simuler $v \sim \mathcal{U}[0, 1]$.
2. Poser $\theta \leftarrow F^-(v)$, où F^- est le pseudo-inverse de F défini par $F^-(v) = \inf\{\theta;\, F(\theta) \geqslant v\}$.

La simulation par méthode d'inversion repose sur un changement de variable qui permet de simuler des variables aléatoires à partir de la loi uniforme. Dans la majorité des cas pratiques l'expression du pseudo-inverse F^- de la fonction de répartition n'est toutefois pas explicite et la méthode n'est pas utilisable directement. D'autres méthodes de simulation basées sur des changements de variables sont présentés par Devroye (1986).

La méthode d'acceptation-rejet (voir Marin and Robert, 2007, par exemple) que nous décrivons en algorithme 1.2 est une méthode générale qui permet de simuler des variables aléatoires distribuées suivant une loi de densité φ, à partir de variables aléatoires distribuées suivant une loi instrumentale q, à la condition que le support de φ soit contenu dans le support de q et que le rapport φ/q soit majoré par $M \geqslant 1$.

Algorithme 1.2 (Méthode d'acceptation-rejet)**.** Pour simuler $\theta \sim \varphi(\theta)$

1. Simuler deux variables indépendantes $\xi \sim q(\xi)$ et $v \sim \mathcal{U}[0,1]$.
2. Calculer
$$\rho(\xi) = \frac{\varphi(\xi)}{Mq(\xi)}.$$
3. Si $v \leqslant \rho(\xi)$ alors poser $\theta \leftarrow \xi$, sinon retourner à la première étape.

La mise en oeuvre de cet algorithme ne requiert pas le calcul explicite de la borne M, puisqu'il est suffisant de connaître les densités φ et q à une constante multiplicative près. Puisque M représente le nombre moyen de simulations à effectuer suivant q pour obtenir une simulation suivant φ, il est important de choisir q avec soin de manière à obtenir un générateur aléatoire le plus efficace possible. Ceci suppose en particulier une connaissance préalable plutôt fine de la loi φ.

Simulation approchée

La simulation directe par des méthodes classiques ne fournissant pas une réponse aux problèmes pratiques rencontrés pour l'inférence, nous présentons ci-dessous l'approche alternative la plus commune (voir par exemple Robert and Casella, 2009, pour des exemples d'utilisations). Cette approche repose sur la simulation de chaînes de Markov $(\theta^t)_{t\in\mathbb{N}}$ de loi stationnaire φ vérifiant un théorème ergodique (voir Robert and Casella, 2004, pour les détails théoriques) i.e. telles que pour tout $h \in \mathrm{L}^1(\varphi)$

$$\frac{1}{M}\sum_{j=1}^{M} h(\theta^j) \xrightarrow{p.s.} \mathbb{E}^{\varphi}[h(\theta)]. \tag{1.18}$$

La définition de l'estimateur (1.15), construit originellement dans le cadre de simulations i.i.d., peut alors s'étendre au cadre de telles chaînes de Markov et vérifie, sous des hypothèses supplémentaires, un TCL semblable à (1.16). Relâcher l'hypothèse d'indépendance sur les variables aléatoires simulées possède toutefois un coût puisqu'il devient alors délicat de construire un estimateur $\widehat{\sigma}_n^2$ de la variance asymptotique similaire à (1.17).

Notons au passage que la fonction de répartition empirique de $\theta^1, \ldots, \theta^M$ converge vers la fonction de répartition associée à φ, et que les quantiles empiriques convergent alors également vers les quantiles de la loi φ (voir van der Vaart, 2000, à nouveau).

Nous présentons brièvement ci-dessous les deux algorithmes les plus populaires qui produisent des chaînes de Markov de loi stationnaire $\varphi(\theta)$, et dont nous faisons notamment usage dans le chapitre 3.

Algorithme de Metropolis-Hastings. Décrit originellement par Metropolis et al. (1953) puis généralisé par Hastings (1970), l'algorithme 1.3 est très général et permet de générer une chaîne de Markov à partir d'une densité instrumentale $q(\xi|\theta)$. La loi $\varphi(\theta)$ est la loi stationnaire de la chaîne de Markov dès que celle-ci est irréductible (voir Robert and Casella, 2004), ce qui est notamment le cas quand le support de $q(\xi|\theta)$ contient le support de $\varphi(\theta)$ pour tout θ (voir par exemple Robert and Casella, 2004, pour cette condition suffisante).

Algorithme 1.3 (Algorithme de Metropolis-Hastings).

```
Pour $t = 0$
        Choisir arbitrairement $\theta^0$.
Pour $t \geqslant 1$
    1. Simuler $\xi \sim q(\xi|\theta^{t-1})$.
    2. Calculer
       $$\rho(\theta^{t-1}, \xi) = \min\left(\frac{\varphi(\xi)q(\theta^{t-1}|\xi)}{\varphi(\theta^{t-1})q(\xi|\theta^{t-1})}, 1\right).$$
    3. Simuler $v \sim \mathcal{U}[0, 1]$. Poser $\theta^t \leftarrow \xi$ si $v \leqslant \rho$, et poser $\theta^t \leftarrow \theta^{t-1}$ sinon.
```

Il est important de souligner que l'exploitation de cet algorithme ne requiert la connaissance de φ qu'à une constante multiplicative près : en particulier, pour $\varphi(\theta) = \pi(\theta|x)$, le calcul de la constante de normalisation n'est pas nécessaire en pratique.

Le choix le plus commun pour la densité instrumentale q consiste à considérer une densité q de la forme $q(\xi|\theta) = \psi(\|\theta - \xi\|)$. Ce choix permet de réécrire la valeur ξ proposée comme une perturbation de l'état précédent de la chaîne de Markov, et présente également l'avantage de simplifier le calcul de la probabilité d'acceptation ρ comme indiqué dans l'algorithme 1.4.

Algorithme 1.4 (Algorithme de Metropolis-Hastings, random-walk proposal).

```
Pour $t = 0$
        Choisir arbitrairement $\theta^0$.
Pour $t \geqslant 1$
    1. Simuler $\epsilon \sim \psi(\epsilon)$ et poser $\xi \leftarrow \theta^{t-1} + \epsilon$.
    2. Calculer
       $$\rho(\theta^{t-1}, \xi) = \min\left(\frac{\varphi(\xi)}{\varphi(\theta^{t-1})}, 1\right).$$
    3. Simuler $v \sim \mathcal{U}[0, 1]$. Poser $\theta^t \leftarrow \xi$ si $v \leqslant \rho$, et poser $\theta^t \leftarrow \theta^{t-1}$ sinon.
```

Echantillonnage de Gibbs. Lorsque les lois conditionnelles $\varphi_1(\theta_1|\theta_2, \dots, \theta_d), \dots, \varphi_1(\theta_d|\theta_1, \dots, \theta_{d-1})$ sont toutes accessibles par simulation, l'échantillonnage de Gibbs (voir Geman and Geman, 1984, pour l'application originale aux champs de Gibbs) représente une solution alternative à l'algorithme de Metropolis-Hastings. Cette méthode, décrite en algorithme 1.5, est basée sur une mise à jour incrémentale des coordonnées de θ grâce aux lois conditionnelles. La chaîne de Markov produite par échantillonnage de Gibbs est ergodique de loi stationnaire φ, sous contrainte de positivité, i.e. dès que le support de φ est le produit cartésien des supports des lois conditionnelles φ_j, pour $j = 1, \dots, d$ (voir la preuve donnée par Robert and Casella, 2004, par exemple).

Algorithme 1.5 (Echantillonnage de Gibbs)**.**

```
Pour $t = 0$
        Choisir arbitrairement $\theta^0 = (\theta_1^0, \dots, \theta_d^0)$.
Pour $t \geqslant 1$
    1. Simuler $\theta_1^t \sim \varphi_1(\theta_1 | \theta_2^{t-1}, \theta_3^{t-1}, \dots, \theta_d^{t-1})$.
    2. Simuler $\theta_2^t \sim \varphi_2(\theta_2 | \theta_1^t, \theta_3^{t-1}, \dots, \theta_d^{t-1})$.
       ⋮
    $d$. Simuler $\theta_d^t \sim \varphi_d(\theta_d | \theta_1^t, \theta_2^t, \dots, \theta_{d-1}^t)$.
```

Signalons que même si l'algorithme 1.5 parait moins générique que l'algorithme 1.3, au sens où il est nécessaire d'identifier les lois φ_j et de pouvoir y accéder par la simulation, il n'en demeure pas moins un choix populaire car il est particulièrement adapté aux modèles hiérarchiques pour lequel les distributions conditionnelles s'avèrent faciles à exprimer.

Echantillonnage d'importance

Nous refermons cet aperçu des méthodes stochastiques pour le calcul des quantités d'intérêt en décrivant rapidement les techniques d'échantillonnage d'importance (importance sampling) sur lesquelles reposent notamment les techniques d'estimations séquentielles présentées en détails dans le Chapitre 4. L'estimateur d'échantillonnage d'importance, semblable à l'estimateur de Monte Carlo (1.15), étend les résultats obtenus pour celui-ci à un cadre plus large dans lequel nous ne simulons plus directement des variables aléatoires suivant la loi d'intérêt mais des variables aléatoires suivant une loi instrumentale.

Notant à nouveau q la loi instrumentale, choisie telle que le support de ϕ soit inclus dans celui de q, nous réécrivons (1.13) sous la forme

$$I = \int h(x)\varphi(x)\,\mathrm{d}x = \int \frac{h(x)\varphi(x)}{q(x)} q(x)\,\mathrm{d}x.$$

Puisque $h \in \mathrm{L}^1(\varphi)$, alors $h\varphi/q \in \mathrm{L}^1(q)$ et il est donc possible d'utiliser l'estimateur de Monte Carlo en fixant q et non plus φ comme loi (fictive) d'intérêt. Rappelons alors que pour $\theta^1, \dots, \theta^M$ des variables i.i.d. distribuées suivant la loi de densité $q(\theta)$, l'estimateur d'échantillonnage d'importance de I est défini par

$$\widehat{I}_M^{\text{IS}}(q) = \frac{1}{M} \sum_{j=1}^{M} \frac{\pi(\theta^j) h(\theta^j)}{q(\theta^j)} = \frac{1}{M} \sum_{j=1}^{M} \widetilde{w}^j h(\theta^j) \tag{1.19}$$

où

$$\widetilde{w}^j = \frac{\varphi(\theta^j)}{q(\theta^j)}. \tag{1.20}$$

Comme pour l'estimateur (1.15) de Monte Carlo classique, il est facile de constater que pour $h \in \mathrm{L}^1(\varphi)$

$$\mathbb{E}^q[\widehat{I}_M^{\mathrm{IS}}(q)] = I, \qquad\qquad \widehat{I}_M^{\mathrm{IS}}(q) \xrightarrow{p.s.} I.$$

La liberté du choix de la loi d'importance confère à la technique d'échantillonnage d'importance une grande versatilité. Ce choix revêt une importance cruciale (voir Robert and Casella, 2009, par exemple) car la variance de l'estimateur (1.19) n'est finie que lorsque

$$\int \frac{h^2(\theta)\varphi^2(\theta)}{q(\theta)}\,\mathrm{d}\theta < +\infty,$$

et vaut alors

$$\int \frac{h^2(\theta)\varphi^2(\theta)}{q(\theta)}\,\mathrm{d}\theta - \left\{\int h(\theta)\varphi(\theta)\,\mathrm{d}\theta\right\}^2.$$

L'intérêt de l'échantillonnage d'importance est alors clair car il peut fournir, pour un choix judicieux de φ, un estimateur dont la variance est finie et inférieure à celle de l'estimateur (1.15) de Monte Carlo. Il est cependant aisé de voir qu'un mauvais choix de q peut conduire à un estimateur de très grande variance.

Si la technique d'échantillonnage d'importance permet d'approcher le calcul d'intégrales de la forme (1.13), elle ne permet toutefois pas d'approcher directement la loi φ et ses quantiles. Pour cela, il est nécessaire de renormaliser les poids et de considérer un échantillonnage d'importance renormalisé, en construisant l'estimateur

$$\widehat{I}_M^{\mathrm{ISr}}(q) = \frac{\displaystyle\sum_{j=1}^{M} \frac{\pi(\theta^j)h(\theta^j)}{q(\theta^j)}}{\displaystyle\sum_{j=1}^{M} \frac{\pi(\theta^j)}{q(\theta^j)}} = \sum_{j=1}^{M} w^j h(\theta^j) \tag{1.21}$$

où les poids normalisés w^j sont définis par

$$w^j = \frac{\widetilde{w}^j}{\sum_{k=1}^{M} \widetilde{w}^k}, \tag{1.22}$$

et où $\widetilde{w}^j$ désignent les poids non normalisés décrits en (1.20). L'estimateur (1.21) est presque sûrement consistant, asympotiquement sans biais (voir Geweke, 1989, par exemple). Puisque la somme des poids w^j vaut 1, la mesure empirique pondérée

$$\mu_M^{\mathrm{ISr}}(\cdot) = \sum_{j=1}^{M} w^j \delta(\theta^j, \cdot) \tag{1.23}$$

où $\delta(\theta, \cdot)$ désigne la masse de Dirac au point θ, définit une mesure de probabilité et elle converge vers la loi d'intérêt φ. Les quantiles empiriques associés à cette mesure μ_M^{ISr} convergent donc à nouveau vers les quantiles de la loi φ (voir van der Vaart, 2000).

Il est également possible d'approcher la loi φ en simulant des variables i.i.d. suivant la densité μ_M^{ISr}, ce qui conduit à la méthode SIR (Sampling Importance Resampling) décrite par Rubin (1988) et dont les estimateurs sont définis de la manière suivante :

$$\mu^{\text{SIR}}_{\widetilde{M}}(\cdot) = \frac{1}{\widetilde{M}} \sum_{j=1}^{\widetilde{M}} \delta(\widetilde{\theta}^j, \cdot), \tag{1.24}$$

$$\widehat{I}^{\text{SIR}}_{\widetilde{M}}(q) = \frac{1}{\widetilde{M}} \sum_{j=1}^{\widetilde{M}} h(\widetilde{\theta}^j), \tag{1.25}$$

où les variables aléatoires $\widetilde{\theta}^j$, $j = 1, \ldots, \widetilde{M}$, sont (conditionnellement à $\theta^1, \ldots, \theta^M$) i.i.d. et distribuées suivant la loi μ^{ISr}_M définie en (1.23). Il est toujours préférable de choisir l'estimateur (1.21) plutôt que (1.25) puisque le premier possède une variance toujours inférieure à la variance du second (les deux estimateurs possédant le même biais).

1.3 LES PROBLÉMATIQUES ABORDÉES

Le présent manuscript rassemble, dans trois chapitres distincts, nos contributions aux problématiques abordées au cours de la thèse. Les chapitres 2, 3 et 4 qui suivent sont rédigés en anglais sous la forme d'articles. Nous présentons brièvement ci-dessous l'objet de chacun de ces chapitres. La section 1.3.1 présente le contenu du chapitre 2 qui s'intéresse au comportement asymptotique des estimateurs de Bayes pour le modèle de part chauffage. La section 1.3.2 décrit la démarche du chapitre 3 dont l'objectif est la construction d'une loi a priori hiérarchique destinée à améliorer les prévisions en situation d'historique court. Enfin la section 1.3.3 résume le chapitre 4 qui traite de l'estimation bayésienne d'un modèle dynamique dans le but de disposer de prévisions en ligne.

1.3.1 Consistance de la loi a posteriori et de l'estimateur du maximum de vraisemblance pour la régression linéaire par morceaux

Le chapitre 2 est publié dans la revue *Electronic Journal of Statistics* (voir Launay et al., 2012a).

Nous nous intéressons au comportement asymptotique des estimateurs du maximum de vraisemblance et bayésien dans le cadre de la régression linéaire par morceaux, servant par exemple à représenter la part chauffage décrite dans le modèle (1.1). Les observations $X_{1:n} = (X_1, \ldots, X_n)$ dépendent d'une variable exogène $t_{1:n} = (t_1, \ldots, t_n)$ via le modèle défini par

$$X_i = \gamma \cdot (t_i - u) \mathbb{1}_{[t_i, +\infty[}(u) + \xi_i, \tag{1.26}$$

pour tout $i = 1, \ldots, n$, et où $(\xi_i)_{i \in \mathbb{N}}$ est une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) suivant la loi $\mathcal{N}(0, \sigma^2)$, de variance σ^2 inconnue. Dans cette situation, le paramètre inconnu que nous cherchons à estimer est $\theta = (\gamma, u, \sigma^2)$.

Les propriétés asymptotiques sont prouvées sous l'hypothèse que la fonction de répartition empirique du régresseur $t_{1:n}$ converge vers une fonction de répartition continuement différentiable quand le nombre d'observations croît. Sous cette même hypothèse, Feder (1975) prouve que la consistance de l'estimateur du maximum de vraisemblance et sa normalité asymptotique uniquement pour les paramètres γ et u.

Dans le cadre de modèles réguliers avec observations i.i.d., Ghosh et al. (2006) s'intéressent aux estimateurs bayésiens, et montrent la normalité asymptotique de la distribution a posteriori (théorème de Bernstein-von Mises). Leurs hypothèses de régularité imposent notamment la différentiabilité de la vraisemblance au troisième ordre. Nous rappelons ci-dessous l'énoncé de ce théorème obtenu pour un modèle univarié.

Théorème 1.1 (Ghosh et al. (2006), Théorème 4.2)**.** *Soit $\pi(\cdot)$ la densité de la loi a priori sur $\theta \in \Theta \subset \mathbb{R}$, continue et positive en θ_0, $k_0 \in \mathbb{N}$ tel que*

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

et notons

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

et $\widetilde{\pi}_n(\cdot|X_{1:n})$ la densité de la loi de t conditionnellement aux observations $X_{1:n}$, alors sous des hypothèses de régularité du modèle, pour tout $0 \leqslant k \leqslant k_0$, quand $n \to +\infty$, nous avons

$$\int_{\mathbb{R}} \|t\|^k \left| \widetilde{\pi}_n(t|X_{1:n}) - (2\pi)^{-\frac{1}{2}} |I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2}t' I(\theta_0) t} \right| \mathrm{d}t \xrightarrow{\mathbb{P}} 0, \tag{1.27}$$

où $I(\cdot)$ désigne l'information de Fisher du modèle et θ_0 la vraie valeur du paramètre.

L'étude asymptotique du modèle de régression linéaire par morceaux (1.26) est rendue difficile par l'impossibilité de différentier la vraisemblance dans un voisinage de l'estimateur du maximum de vraisemblance.

Pour étendre les résultats de Ghosh et al. (2006), nous complétons tout d'abord les résultats de Feder (1975) : nous nous assurons dans un premier temps de la consistance de l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ vers la vraie valeur θ_0 et contrôlons la vitesse de convergence associée. Nous rappelons ci-dessous le résultat prouvé.

Théorème 1.2 (Chapitre 2, Théorème 2.9)**.** *Sous les hypothèses A1–A4 présentées dans le chapitre 2, quand $n \to +\infty$, nous avons*

$$n^{\frac{1}{2}} \left(\widehat{\theta}_n - \theta_0 \right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_0)^{-1}\right),$$

où $I(\cdot)$ désigne la matrice définie en (2.7) *et θ_0 la vraie valeur du paramètre.*

Nous définissons ensuite un pseudo-problème, outil théorique original introduit par Sylwester (1965), en retirant formellement les observations situées dans un voisinage de θ_0, de façon à ignorer une fraction asymptotiquement négligeable des observations. Choisir de manière adéquate la vitesse de réduction de ce voisinage permet de garantir la dérivabilité de la vraisemblance dans un voisinage de l'estimateur du maximum de vraisemblance pour le pseudo-problème. Nous utilisons alors les techniques de Ghosh et al. (2006), pour recouvrir, sous des hypothèses similaires, le théorème de Bernstein-von Mises relatif à la distribution a posteriori, puis montrons que le pseudo-problème et le problème original sont asymptotiquement équivalents, étendant ainsi naturellement les résultats obtenus pour le premier au second. Nous rappelons ci-dessous le résultat principal prouvé et son corollaire immédiat.

Théorème 1.3 (Chapitre 2, Théorème 2.2)**.** *Soit $\pi(\cdot)$ la densité de la loi a priori sur $\theta \in \Theta \in \mathbb{R}^d$, continue et positive en θ_0, $k_0 \in \mathbb{N}$ tel que*

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

et notons

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

et $\widetilde{\pi}_n(\cdot|X_{1:n})$ *la densité de la loi a posteriori de t conditionnellement aux observations* $X_{1:n}$*, alors sous les hypothèses A1–A4 présentées en page 28, pour tout* $0 \leqslant k \leqslant k_0$*, quand* $n \to +\infty$*, nous avons*

$$\int_{\mathbb{R}^3} \|t\|^k \left|\widetilde{\pi}_n(t|X_{1:n}) - (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2}t' I(\theta_0) t}\right| \mathrm{d}t \xrightarrow{\mathbb{P}} 0, \tag{1.28}$$

où $I(\cdot)$ *désigne la matrice définie en* (2.7) *et* θ_0 *la vraie valeur du paramètre.*

Corollaire 1.4 (Chapitre 2, Corollaire 2.4)**.** *Soit* $\pi(\cdot)$ *la densité de la loi a priori sur* $\theta \in \Theta \in \mathbb{R}^d$*, continue et positive en* θ_0*, telle que*

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

et notons

$$\widetilde{\theta}_n = \int_{\Theta} \theta \pi_n(\theta|X_{1:n})\, \mathrm{d}\theta,$$

l'estimateur de Bayes de θ*. Alors sous les hypothèses A1–A4 présentées en page 28, quand* $n \to +\infty$*, nous avons*

$$n^{\frac{1}{2}}(\widetilde{\theta}_n - \widehat{\theta}_n) \xrightarrow{\mathbb{P}} 0.$$

Les vitesses de convergence obtenues pour ce modèle non i.d.d. sont cohérentes au sens où elles sont identiques aux vitesses obtenues pour l'estimateur du maximum de vraisemblance pour un modèle i.i.d. dont la vraisemblance continue possède un unique point singulier (voir par exemple Dacunha-Castelle, 1978) ou également pour un modèle de régression dont le régresseur vérifie l'hypothèse A1.

1.3.2 Prévision en situation d'historique court

Le chapitre 3 fait actuellement l'objet d'une soumission.

Dans ce chapitre nous développons la méthodologie bayésienne permettant de prévoir la consommation d'électricité, à partir d'un modèle dérivé du modèle (1.1), pour un jeu de données dont l'historique est court. En situation d'historique court, l'estimation des paramètres du modèle (1.1) par la méthode du maximum de vraisemblance est rendue très délicate à cause du nombre important de paramètres en jeu. Le modèle est généralement surajusté aux données disponibles et fournit par conséquent des prévisions de mauvaise qualité.

Pour améliorer la qualité des prévisions, nous supposons qu'il existe un autre jeu de données pour lequel un historique plus long est disponible (permettant une estimation correcte des paramètres du modèle) et supposons que les deux jeux de données se ressemblent. Plus formellement, notons $\mathcal{A}$ le jeu de données dont l'historique est long, $\mathcal{B}$ le jeu de données dont l'historique est court, $\theta^{\mathcal{A}}, \theta^{\mathcal{B}}$ les paramètres d'intérêt respectifs et $X^{\mathcal{A}}, X^{\mathcal{B}}$ les observations respectives. Nous supposons que les données $X^{\mathcal{A}}$ et $X^{\mathcal{B}}$ sont issues d'un même modèle $f(x|\theta)$ pour des valeurs de paramètres $\theta^{\mathcal{A}}$ et $\theta^{\mathcal{B}}$ assez proches.

Afin d'estimer le paramètre $\theta^{\mathcal{B}}$ nous proposons la construction d'une loi a priori hiérarchique basée sur la loi a posteriori $\pi(\theta|X^{\mathcal{A}})$. L'estimation de $\theta^{\mathcal{B}}$ s'effectue en deux temps :

1. à partir d'une loi a priori non-informative, nous estimons la loi a posteriori correspondant au modèle sur le jeu de données $\mathcal{A}$. Les données $X^{\mathcal{A}}$ étant nombreuses, le choix de la loi a priori (pour peu qu'elle possède une grande variance) n'a qu'une influence mineure sur la loi a posteriori.
2. à partir d'une loi a priori hiérarchique normale prenant en compte $\mu^{\mathcal{A}}, \Sigma^{\mathcal{A}}$ (moyenne et variance de la loi a posteriori $\pi(\theta|X^{\mathcal{A}})$ sur le jeu de données $\mathcal{A}$) et la ressemblance entre les jeux de données, nous estimons la loi a posteriori correspondant au modèle sur le jeu de données $\mathcal{B}$. La ressemblance entre les deux jeux de données est modélisée à l'aide d'un opérateur linéaire diagonal K dont les coefficients sont inconnus, mais a priori centrés autour de la valeur 1. La loi a priori que nous utilisons pour estimer le modèle sur le jeu de données $\mathcal{B}$ est de la forme

$$\theta|K \sim \mathcal{N}(K_1\mu^{\mathcal{A}}, K_2\Sigma^{\mathcal{A}}).$$

Nous donnons une représentons schématique de la méthode d'estimation du modèle sur le jeu de données $\mathcal{B}$ en figure 1.4.

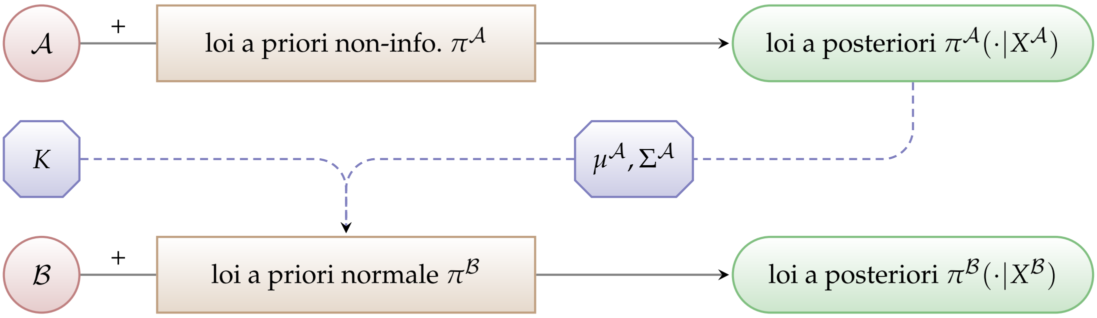

$\mathcal{A}$
+
loi a priori non-info. $\pi^{\mathcal{A}}$
loi a posteriori $\pi^{\mathcal{A}}(\cdot|X^{\mathcal{A}})$
K
$\mu^{\mathcal{A}}, \Sigma^{\mathcal{A}}$
$\mathcal{B}$
+
loi a priori normale $\pi^{\mathcal{B}}$
loi a posteriori $\pi^{\mathcal{B}}(\cdot|X^{\mathcal{B}})$

Figure 1.4: Représentation schématique de la méthode proposée pour estimer les paramètres du modèle en situation d'historique court.

La méthode que nous proposons est tout d'abord validée sur des données que nous simulons : les jeux de données simulées $\mathcal{A}$ et $\mathcal{B}$ sont de longueurs respectives 4 ans et 1 ans, $\theta^{\mathcal{A}}$ est choisi de manière à représenter un comportement de consommation typique au périmètre national, et $\theta^{\mathcal{B}}$ est choisi de la forme $\theta^{\mathcal{B}} = \Lambda\theta^{\mathcal{A}}$ avec Λ opérateur linéaire diagonal connu, pour différents choix de Λ. Par exemple, lorsque seul le seuil de chauffage diffère entre $\mathcal{A}$ et $\mathcal{B}$, i.e. $u^{\mathcal{B}} = \lambda u^{\mathcal{A}}$, nous obtenons les résultats présentés en figures 1.5, 1.6 et 1.7

Les figures 1.5 et 1.6 représentent les erreurs d'estimations sur les paramètres pour $\lambda = 1$ et $\lambda = 0.5$. Dans la situation idéale de ressemblance parfaite ($\lambda = 1$) entre les deux jeux de données $\mathcal{A}$ et $\mathcal{B}$ la méthode informative développée fournit des estimations plus précises qu'une méthode non informative. Quand la ressemblance entre les jeux de données n'est plus exacte ($\lambda = 0.5$) l'estimation des paramètres reste néanmoins correcte : la loi a priori ne déteriore pas la qualité des estimations, par rapport à une méthode non informative.

La figure 1.7 représente le rapport entre le RMSE en prévision obtenu par la méthode informative d'estimation que nous proposons et une méthode non informative (assimilable à une estimation par maximisation de la vraisemblance) en fonction du coefficient λ de ressemblance entre les seuils de chauffage des deux jeux de données $\mathcal{A}$ et $\mathcal{B}$ simulés. Il est clair que la méthode développée améliore

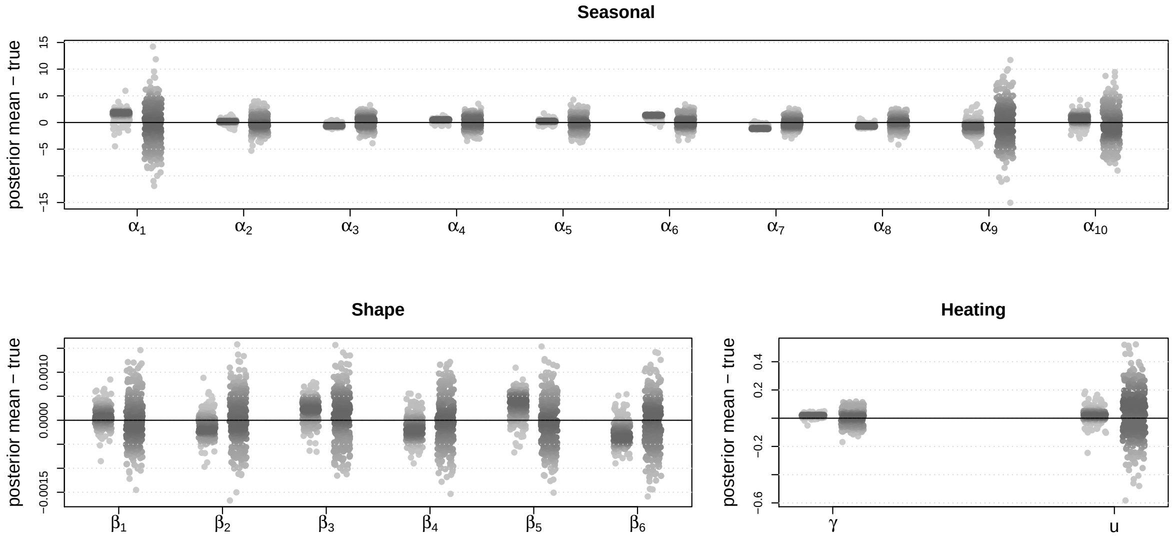


Figure 1.5: Erreur d'estimation a posteriori (différence entre la moyenne a posteriori et la vraie valeur) sur les paramètres de saisonnalité, de formes de jours et de chauffage (gradient et seuil) de $\mathcal{B}$, pour 300 réplications. Les réplications les plus à gauche correspondent à l'utilisation de la loi a priori hiérarchique développée, tandis que les réplications les plus à droite correspondent à l'utilisation d'une loi a priori non informative. Ici $u^{\mathcal{B}} = \lambda u^{\mathcal{A}}$, avec $\lambda = 1$.

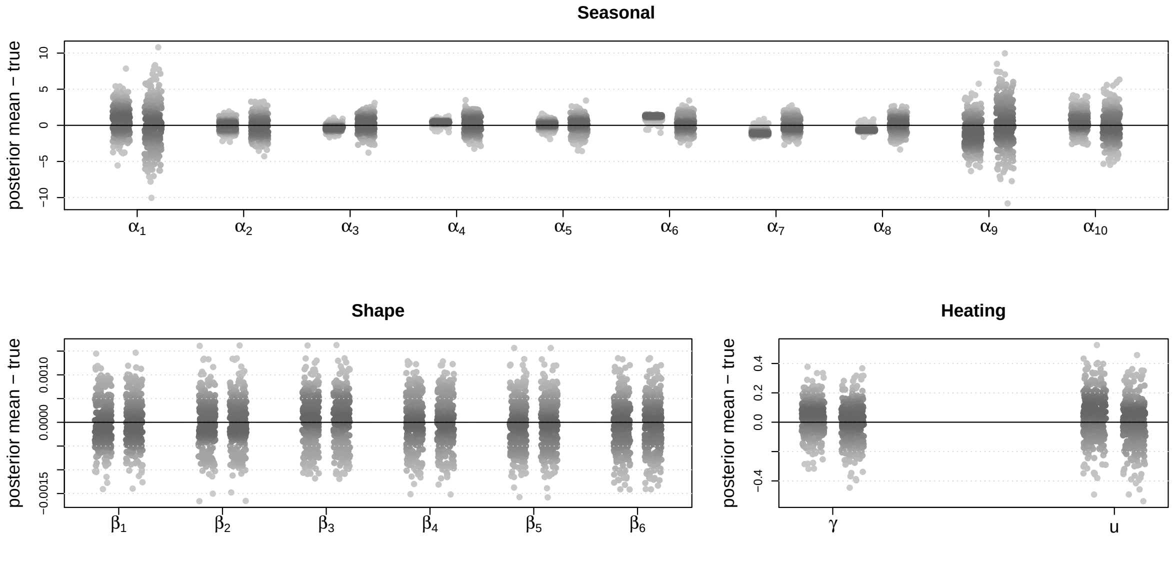


Figure 1.6: Même légende qu'en figure 1.5 pour $u^{\mathcal{B}} = \lambda u^{\mathcal{A}}$, avec $\lambda = 0.5$.

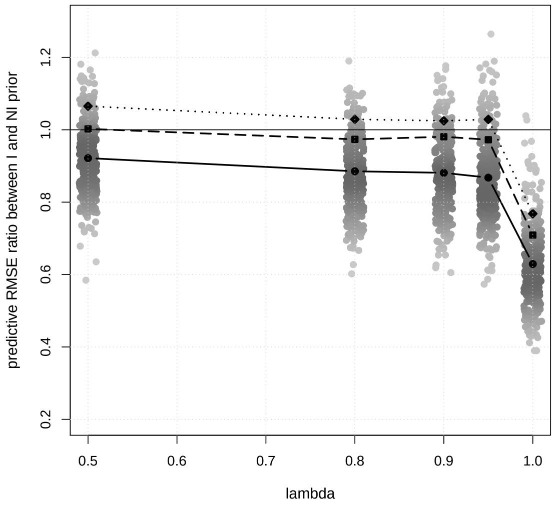


Figure 1.7: Rapports entre le RMSE en prévision sur 365 jours de la méthode informative proposée et celui d'une méthode non informative, en fonction du coefficient λ de ressemblance entre les seuils de chauffage des jeux de données simulés. Les points en nuances de gris correspondent à 300 réplications pour chaque valeur de λ testée. Les points en noir indiquent les moyennes obtenues et les carrés et losanges représentent les quantiles empiriques à 80% and 90% de ces rapports.

substantiellement la qualité des prévisions quand la ressemblance entre les jeux de données est parfaite ($\lambda = 1$, avec un gain approchant 40% en moyenne), mais également quand ce n'est plus le cas (avec un gain approchant 10% en moyenne).

Nous appliquons enfin notre méthode sur des jeux de données réels et comparons les performances obtenues en prévision avec d'autres modèles. Nous montrons notamment, en réduisant progressivement la longueur de l'historique de données considéré, que la loi a priori informative que nous proposons permet de rendre l'estimation du modèle plus robuste vis-à-vis du manque de données. La table 1.1 présente les erreurs d'estimation et de prévision obtenues en fonction de la longueur de l'historique. Des résultats similaires, pour d'autres situations de ressemblance sont présentés dans le chapitre 3.

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

Table 1.1: Qualité globale (RMSE en MW, et MAPE en %) pour l'estimation (gauche) et la prévison (droite) à partir des lois a priori non informative (non info.) et informative (info.), en fonction du nombre de mois utilisés pour la période d'estimation (de 12 mois à 6 mois). La population $\mathcal{A}$ utilisée correspond aux profilés EDF (clients non télérelevés), la population $\mathcal{B}$ utilisée correspond aux profilés ERDF.

1.3.3 Application du filtrage particulaire à la prévision de consommation d'électricité

Dans le chapitre 4 nous nous intéressons à la prévision en ligne de la consommation d'électricité à partir d'un modèle à espace d'états, i.e. dont les paramètres varient au cours du temps.

Soit $\{X_n\}_{n\geqslant 0}$ et $\{Y_n\}_{n\geqslant 0}$ des processus aléatoires à valeurs dans $\mathcal{X} \subset \mathbb{R}^{n_x}$ et $\mathcal{Y} \subset \mathbb{R}^{n_y}$ et définis sur un espace mesurable. Nous supposons les observations $\{Y_n\}_{n\geqslant 0}$ indépendantes conditionellement au processus de Markov caché $\{X_n\}_{n\geqslant 0}$, appelé état du modèle, et caractérisées par la loi conditionnelle $g_n^\theta(y_n|x_n)$. Nous notons $\mu^\theta(x_0)$ la densité initiale de l'état et $f_n^\theta(x_n|x_{n-1})$ la densité de la transition de Markov du temps $n-1$ au temps n. L'indice θ de ses densités correspond au paramètre du modèle, et appartient à un ensemble ouvert $\Theta \subset \mathbb{R}^{n_\theta}$. Nous pouvons résumer le modèle sous l'écriture synthétique

$$X_0 \sim \mu^\theta(\cdot), \quad X_n|(X_{n-1} = x_{n-1}) \sim f_n^\theta(\cdot|x_{n-1}) \tag{1.29}$$

$$Y_n|(X_n = x_n) \sim g_n^\theta(\cdot|x_n). \tag{1.30}$$

Dans le contexte bayésien décrit en section 1.2, l'équation (1.29) spécifie la loi a priori sur l'état du modèle dont la vraisemblance est définie via (1.30).

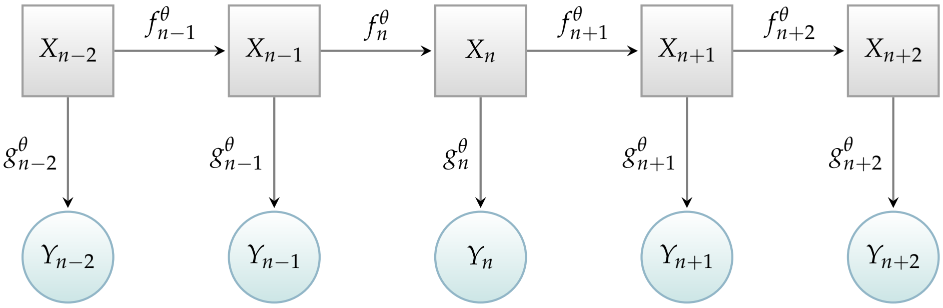


Figure 1.8: Représentation schématique d'un modèle de Markov caché.

Nous nous intéressons ici uniquement à des modèles dont les observations sont indépendantes, mais il est aisé d'étendre le cadre de travail à des obsevations dépendantes. Les modèles dynamiques que nous considérons sont appelés modèles à espace d'états ou modèle de Markov caché dans la littérature et leur représentation générique est fournie en figure 1.8. Cette classe de modèles inclut en particulier de nombreux modèles non linéaires et non gaussiens de séries temporelles.

Notre principal objectif étant d'estimer successivement les lois a posteriori

$$\pi^\theta(x_{0:n}|y_{0:n}) \propto \underbrace{\prod_{k=1}^{n} g_k^\theta(y_k|x_k)}_{n \text{ vraisemblances}} \cdot \underbrace{\prod_{k=1}^{n} f_k^\theta(x_k|x_{k-1})}_{n \text{ densités de transition}} \cdot \underbrace{\mu^\theta(x_0)}_{\text{densité initiale}} , \tag{1.31}$$

nous présentons les techniques séquentielles de Monte Carlo, définies à partir des outils décrits en section 1.2.3 et aussi connues sous le nom de filtres particulaires. Ces techniques permettent d'estimer les lois marginales $\pi^\theta(x_n|y_{0:n})$ par la mesure empirique pondérée de M variables aléatoires (particules)

X^i pondérées par des poids w^i, $i = 1, \dots, M$ i.e.

$$\sum_{j=1}^{M} w^j \delta(X^j, \cdot)$$

Les deux étapes essentielles au filtrage particulaire, dérivées de la méthode SIR introduite en section 1.2.3, sont schématisées en figure 1.9 et reposent respectivement sur les équations (1.29) et (1.30) qui définissent le modèle dynamique :

1. à partir de M variables aléatoires (particules) X^i_{n-1} pondérées par des poids w^i_{n-1} représentant la loi filtrée $\pi^\theta(x_{n-1}|y_{0:n-1})$ au temps $n-1$, la première étape consiste à simuler, à l'aide de la densité de transition $f^\theta_n(x_n|x_{n-1})$ (voir équation (1.29)), de nouvelles particules $X^i_{n|n-1}$ pondérées par des poids $w^i_{n|n-1}$ pour approcher la loi prédictive du prochain état $\pi^\theta(x_n|y_{0:n-1})$;
2. le rôle de la seconde étape est de corriger l'échantillon de particules pondérées ainsi simulées, grâce à la vraisemblance $g^\theta_n(y_n|x_n)$ de la donnée y_n observée (voir équation (1.30)), pour obtenir des particules X^i_n pondérées par des poids w^i_n pour représenter la loi filtrée $\pi^\theta(x_n|y_{0:n})$ au temps n.

Nous discutons les deux principaux problèmes qui surviennent lors de l'utilisation de filtres particulaires à savoir : la dégénérescence asymptotique de la distribution des particules et l'estimation conjointe du paramètre θ (lorsque celui-ci n'est pas connu).

Nous décrivons les algorithmes détaillés associés à des solutions parmi les moins coûteuses (en terme de calculs) pour une implémentation pratique la plus directe possible. Enfin, nous présentons l'algorithme que nous retenons et utilisons pour l'estimation de modèles dynamiques de consommation d'électricité. L'originalité de cet algorithme repose sur la détection et la suppression automatique des données aberrantes qui conduise en temps normal à la dégénerescence de la distribution des particules.

Le modèle dynamique que nous proposons est basé sur le modèle (1.1) au sens où il repose sur la définition de trois parts : saisonnalité, chauffage et climatisation, les deux premières étant choisies dynamiques (avec par exemple un coefficient de saisonnalité et un gradient de chauffage variant au cours du temps), et la dernière fixe. La composante chauffage est modélisée par

$$x_n^{\text{heat}} = g_n^{\text{heat}}(T_n^{\text{heat}} - u^{\text{heat}})\mathbb{1}_{]T_n^{\text{heat}},+\infty[}(u^{\text{heat}})$$

où u est un paramètre inconnu constant à estimer et où g_n évolue au cours du temps suivant une dynamique de marche aléatoire

$$g_n^{\text{heat}} = g_{n-1}^{\text{heat}} + \epsilon_n^g.$$

La composante saisonnalité est quant à elle modélisée simplement par

$$x_n^{\text{season}} = s_n \cdot \kappa_{\text{daytype}_n}$$

où les coefficients κ_k sont des paramètres constants à estimer (permettant de modéliser l'effet type de jour), et où s_n évolue au cours du temps suivant une dynamique de marche aléatoire

$$s_n = s_{n-1} + \epsilon_n^s.$$

Le modèle étudié (voir la définition complète donnée au chapitre 4) est différent du modèle proposé par Dordonnat et al. (2008) car

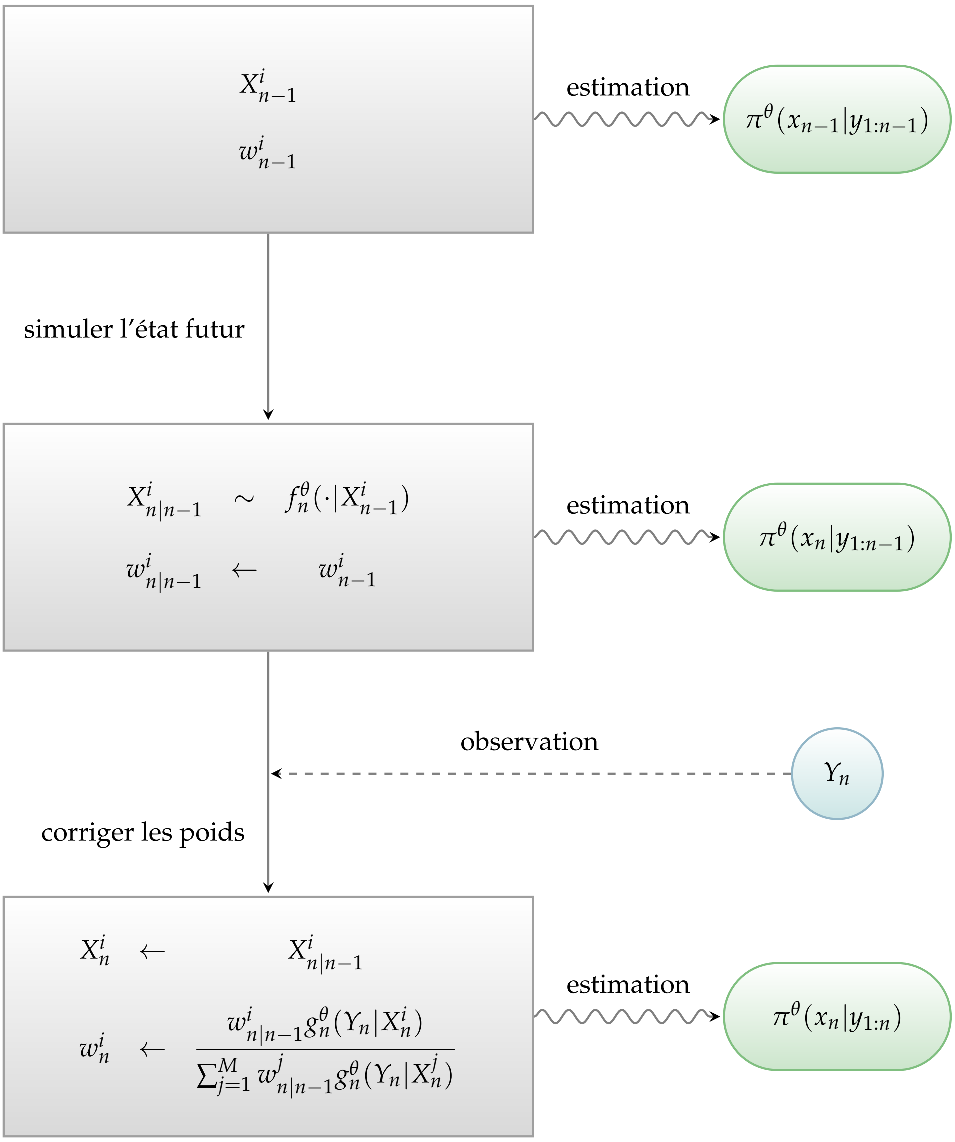

X^i_{n-1}
w^i_{n-1}
estimation
$\pi^\theta(x_{n-1}|y_{1:n-1})$
simuler l'état futur
$X^i_{n|n-1} \sim f^\theta_n(\cdot|X^i_{n-1})$
$w^i_{n|n-1} \leftarrow w^i_{n-1}$
estimation
$\pi^\theta(x_n|y_{1:n-1})$
observation
Y_n
corriger les poids
$X^i_n \leftarrow X^i_{n|n-1}$
$w^i_n \leftarrow \frac{w^i_{n|n-1} g^\theta_n(Y_n|X^i_n)}{\sum_{j=1}^M w^j_{n|n-1} g^\theta_n(Y_n|X^j_n)}$
estimation
$\pi^\theta(x_n|y_{1:n})$

Figure 1.9: Représentation schématique de l'évolution des particules X^i pondérées par des poids w^i, pour $i = 1, \ldots, M$, pour la mise à jour du temps $n-1$ au temps n. A chaque étape, les particules (à gauche) sont utilisées pour approcher différentes lois d'intérêt (à droite) par la mesure empirique pondérée $\sum_{j=1}^M w^j \delta(X^j, \cdot)$.

1. il n'est pas linéaire ;
2. il modélise la saisonnalité de façon plus parcimonieuse ;
3. il autorise la dynamique des coefficients à évoluer elle-même au cours du temps.

Nous montrons à travers différents résultats que le filtre particulaire que nous élaborons permet d'estimer le modèle de manière satisfaisante, et répond à nos exigences de robustesse vis-à-vis de données aberrantes. Nous examinons également la qualité des prévisions obtenues par le modèle dynamique à partir de différents critères. Par exemple, la figure 1.10 représente les erreurs relatives de prévision du modèle dynamique au cours de la journée et montre que le modèle dynamique fournit des prévisions compétitives avec les prévisions opérationnelles. Enfin, nous comparons notre approche au modèle développé par Dordonnat et al. (2008) et estimé par filtrage de Kalman, en soulignant que les qualités de prévisions obtenues sont similaires malgré le fait que le modèle soit de moindre dimension.

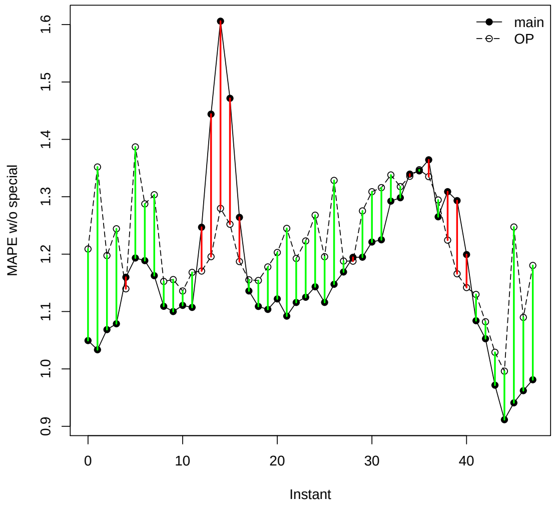


Figure 1.10: MAPE (en %) prédictif, hors jours spéciaux (jours fériés, etc.), pour le modèle dynamique (4.23) défini au chapitre 4 et le modèle opérationnel pour chacun des 48 instants de la journée. La différence entre les deux modèles est colorée en fonction de son signe : en vert quand le modèle dynamique est meilleur que le modèle opérationnel, en rouge sinon.

2 Consistency of the posterior distribution and MLE for piecewise linear regression

We prove the weak consistency of the posterior distribution and that of the Bayes estimator for a two-phase piecewise linear regression model where the break-point is unknown. We also establish a Bernstein-von Mises theorem for this non regular model. The non differentiability of the likelihood of the model with regard to the break-point parameter induces technical difficulties that we overcome by creating a regularised version of the problem at hand. We first recover the strong consistency of the quantities of interest for the regularised version, using results about the MLE, and we then prove that the regularised version and the original version of the problem share the same asymptotic properties.

2.1 Introduction

We consider a continuous segmented regression model with 2 phases, one of them (the rightmost) being zero. Let u be the unknown breakpoint and $\gamma \in \mathbb{R}$ be the unknown regression coefficient of the non zero phase. The observations $X_{1:n} = (X_1, \ldots, X_n)$ depend on an exogenous variable that we denote $t_{1:n} = (t_1, \ldots, t_n)$ via the model given for $i = 1, \ldots, n$ by

$$X_i = \mu(\eta, t_i) + \xi_i := \gamma \cdot (t_i - u) \mathbb{1}_{[t_i, +\infty[}(u) + \xi_i, \tag{2.1}$$

where $(\xi_i)_{i \in \mathbb{N}}$ is a sequence of independent and identically distributed (i.i.d.) random variables with a common centred Gaussian distribution of unknown variance σ^2, $\mathcal{N}(0, \sigma^2)$, and where $\mathbb{1}_A$ denotes the indicator function of a set A.

Such a model is for instance used in practise to estimate and predict the heating part of the electricity demand in France. See Bruhns et al. (2005) for the definition of the complete model and Launay et al. (2012b) for a Bayesian approach. In this particular case, u corresponds to the heating threshold above which the temperatures $t_{1:n}$ do not have any effect over the electricity load, and γ corresponds to the heating gradient i.e. the strength of the described heating effect.

The work presented in this paper is most notably inspired by the results developed in Ghosh et al. (2006) and Feder (1975).

Feder proved the weak consistency of the least squares estimator in segmented regression problems with a known finite number of phases under the hypotheses of his Theorem 3.10 and some additional assumptions disseminated throughout his paper, amongst which we find that the empirical cumulative distribution functions of the temperatures at the n-th step $t_{n1}, \ldots, t_{nn}$ are required to converge to a cumulative distribution function, say F_n converges to F, which is of course to be compared to our own Assumption A1. Feder also derived the asymptotic distribution of the least squares estimator under the same set of assumptions. Unfortunately there are a few typographical errors in his paper (most notably resulting in the disappearance of σ_0^2 from the asymptotic variance matrix in his main theorems), and he also did not include $\widehat{\sigma}_n^2$ in his study of the asymptotic distribution.

The asymptotic behaviour of the posterior distribution is a central question that has already been raised in the past. For example, Ghosh et al. worked out the limit of the posterior distribution in a general and regular enough i.i.d. setup. In particular they manage to derive the asymptotic normality of the posterior distribution under third-order differentiability conditions. There are also a number of works dealing with some kind of non regularity, like these of Sareen (2003) which consider data the support of which depends on the parameters to be estimated, or those of Ibragimov and Khasminskii (1981) which offer the limiting behaviour of the likelihood ratio for a wide range of i.i.d. models whose likelihood may present different types of singularity. Unfortunately, the heating part model presented here does not fall into any of these already studied categories.

In this paper, we show that the results of Ghosh et al. can be extended to a non i.i.d. two-phase regression model. We do so by using the original idea found in Sylwester (1965) [1]: we introduce a new, regularised version of the problem called pseudo-problem, later reprised by Feder. The pseudo-problem consists in removing a fraction of the observations in the neighbourhood of the true parameter to obtain a differentiable likelihood function. We first recover the results of Ghosh et al. for this pseudo-problem and then extend these results to the (full) problem by showing that the estimates for the problem and the pseudo-problem have the same asymptotic behaviour.

From this point on, we shall denote the parameters $\theta = (\gamma, u, \sigma^2) = (\eta, \sigma^2)$ and θ_0 will denote the true value of θ. We may also occasionally refer to the intercept of the model as $\beta = -\gamma u$. The log-likelihood of the n first observations $X_{1:n}$ of the model will be denoted

$$\begin{aligned} l_{1:n}(X_{1:n}|\theta) &= \sum_{i=1}^{n} l_i(X_i|\theta) && (2.2) \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n}\frac{1}{2\sigma^2}\left(X_i - \gamma\cdot(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u)\right)^2, && (2.3) \end{aligned}$$

where $l_i(X_i|\theta)$ designates the log-likelihood of the i-th observation X_i, i.e.

$$l_i(X_{1:n}|\theta) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(X_i - \gamma\cdot(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u)\right)^2. \quad (2.4)$$

Notice that we do not mention explicitly the link between the likelihood l and the sequence of temperatures $(t_n)_{n\in\mathbb{N}}$ in these notations, so as to keep them as minimal as possible. The least square estimator $\widehat{\theta}_n$ of θ being also the maximum likelihood estimator of the model, we refer to it as the MLE.

Throughout the rest of this paper we work under the following assumptions

Assumption A1. The sequence of temperatures (exogenous variable) $(t_n)_{n\in\mathbb{N}}$ belongs to a compact set $[\underline{u}, \overline{u}]$ and the sequence of the empirical cumulative distribution functions $(F_n)_{n\in\mathbb{N}}$ of $(t_1, \ldots, t_n)$,

[1] Sylwester indeed considers the same model as we do here, however his asymptotic results are false due to an incorrect reparametrisation of the problem and an error in the proof of his Theorem 3.5.

defined by

$$F_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[t_i,+\infty[}(u),$$

converges pointwise to a function F where F is a cumulative distribution function itself, which is continuously differentiable over $[\underline{u}, \overline{u}]$.

Remark 1. Due to a counterpart to Dini's Theorem (see Theorem 2.11 taken from Polya and Szegö, 2004, p81), F_n converges to F uniformly over $[\underline{u}, \overline{u}]$.

Remark 2. Let h be a continuous, bounded function on $[\underline{u}, \overline{u}]$. As an immediate consequence of this assumption, for any interval $I \subset [\underline{u}, \overline{u}]$, we have, as $n \to +\infty$

$$\frac{1}{n}\sum_{i=1}^{n} h(t_i)\mathbb{1}_I(t_i) = \int_I h(t)\,\mathrm{d}F_n(t) \to \int_I h(t)\,\mathrm{d}F(t) = \int_I h(t)f(t)\,\mathrm{d}t,$$

the convergence holding true by definition of the convergence of probability measures (see Billingsley, 1999, pages 14–16). In particular, for $I = [\underline{u}, \overline{u}]$ and $I =]-\infty, u]$ we get, as $n \to +\infty$

$$\frac{1}{n}\sum_{i=1}^{n} h(t_i) \to \int_{\underline{u}}^{\overline{u}} h(t)f(t)\,\mathrm{d}t, \qquad \frac{1}{n}\sum_{i=1}^{n} h(t_i)\mathbb{1}_{[t_i,+\infty[}(u) \to \int_{\underline{u}}^{u} h(t)f(t)\,\mathrm{d}t.$$

Remark 3. It is a general enough assumption which encompasses both the common cases of i.i.d. continuous random variables and periodic (non random) variables under a continous (e.g. Gaussian) noise.

Assumption A2. $\theta_0 \in \Theta$, where the parameter space Θ is defined (for identifiability) as

$$\Theta = \mathbb{R}^* \times]\underline{u}, \overline{u}[\times \mathbb{R}_+^*,$$

where $\mathbb{R}^* = \{x \in \mathbb{R}\,, x \neq 0\}$ and $\mathbb{R}_+^* = \{x \in \mathbb{R}\,, x > 0\}$.

Assumption A3. $f = F'$ does not vanish (i.e. is positive) on $]\underline{u}, \overline{u}[$.

Assumption A4. There exists $K \subset \Theta$ a compact subset of the parameter space Θ such that $\widehat{\theta}_n \in K$ for any n large enough.

The paper is organised as follows. In Section 2.2, we present the Bayesian consistency (the proofs involved there rely on the asymptotic distribution of the MLE) and introduce the concept of pseudo-problem. In Section 2.3, we prove that the MLE for the full problem is strongly consistent. In Section 2.4 we derive the asymptotic distribution of the MLE using the results of Section 2.3: to do so, we first derive the asymptotic distribution of the MLE for the pseudo-problem and then show that the MLEs for the pseudo-problem and the problem share the same asymptotic distribution. We discuss these results in Section 2.5. The extensive proofs of the main results are found in Section 2.6 while the most technical results are pushed back into Section 2.7 at the end of this paper.

Notations. Whenever mentioned, the O and o notations will be used to designate a.s. O and a.s. o respectively, unless there are indexed with $\mathbb{P}$ as in $\mathrm{O}_{\mathbb{P}}$ and $\mathrm{o}_{\mathbb{P}}$, in which case they will designate O and o in probability respectively.

Hereafter we will use the notation A^c for the complement of the set A and $B(x, r)$ for the open ball of radius r centred at x i.e. $B(x, r) = \{x',\ \|x' - x\| < r\}$.

2.2 BAYESIAN CONSISTENCY

In this Section, we show that the posterior distribution of θ given $(X_1, \dots, X_n)$ asymptotically favours any neighbourhood of θ_0 as long as the prior distribution itself charges a (possibly different) neighbourhood of θ_0 (see Theorem 2.1). We then present in Theorem 2.2 the main result of this paper i.e. the convergence of posterior distribution with suitable normalisation to a Gaussian distribution.

2.2.1 Consistency and asymptotic normality of the posterior distribution

Theorem 2.1. *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 A1–A4, as $n \longrightarrow +\infty$,*

$$\int_U \pi(\theta|X_{1:n})\,\mathrm{d}\theta \xrightarrow{a.s} 1. \tag{2.5}$$

Proof for Theorem 2.1. The proof is very similar to the one given in Ghosh and Ramamoorthi (2003) for a model with i.i.d. observations. Let $\delta > 0$ small enough so that $B(\theta_0, \delta) \subset U$. Since

$$\begin{aligned}\int_U \pi(\theta|X_{1:n})\,\mathrm{d}\theta &= \cfrac{1}{1+\cfrac{\int_{U^c} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta}{\int_U \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta}} \\ &\geqslant \cfrac{1}{1+\cfrac{\int_{B^c(\theta_0,\delta)} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta}{\int_{B(\theta_0,\delta)} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta}}\end{aligned}$$

it will suffice to show that

$$\frac{\int_{B^c(\theta_0,\delta)} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta}{\int_{B(\theta_0,\delta)} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta} \xrightarrow{a.s} 0. \tag{2.6}$$

To prove (2.6) we adequately majorate its numerator and minorate its denominator. The majoration mainly relies on Proposition 2.21 while the minoration is derived without any major difficulties. The comprehensive proof of (2.6) can be found in Section 2.6.1 on page 38. ■

Let $\theta \in \Theta$, we now define $I(\theta)$, the asymptotic Fisher Information matrix $I(\theta)$ of the model, as the symmetric matrix given by

$$I(\theta) = \begin{bmatrix} \sigma^{-2}\int_{\underline{u}}^{u}(t-u)^2\,\mathrm{d}F(t) & -\sigma^{-2}\gamma\int_{\underline{u}}^{u}(t-u)\,\mathrm{d}F(t) & 0 \\ & \sigma^{-2}\gamma^2\int_{\underline{u}}^{u} 1\,\mathrm{d}F(t) & 0 \\ & & \frac{1}{2}\sigma^{-4} \end{bmatrix}. \tag{2.7}$$

It is obviously positive and definite since all its principal minor determinants are positive. The proof of the fact that it is indeed the limiting matrix of the Fisher Information matrix of the model is deferred to Lemma 2.20.

Theorem 2.2. *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

$$t = n^{\frac{1}{2}}(\theta - \widehat{\theta}_n), \tag{2.8}$$

and $\widetilde{\pi}_n(\cdot|X_{1:n})$ *the posterior density of t given* $X_{1:n}$*, then under Assumptions A1–A4, for any* $0 \leqslant k \leqslant k_0$*, as* $n \to +\infty$*,*

$$\int_{\mathbb{R}^3} \|t\|^k \left|\widetilde{\pi}_n(t|X_{1:n}) - (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2}t'I(\theta_0)t}\right| \mathrm{d}t \xrightarrow{\mathbb{P}} 0, \tag{2.9}$$

where $I(\theta)$ *is defined in* (2.7) *and* θ_0 *the true value of the parameter.*

The proof Theorem 2.2 relies on the consistency of the pseudo-problem, first introduced in Sylwester (1965), that we define in the next few paragraphs.

2.2.2 Pseudo-problem

The major challenge in proving Theorem 2.2 is that the typical arguments usually used to derive the asymptotic behaviour of the posterior distribution (see Ghosh et al., 2006, for example) do not directly apply here. The proof provided by Ghosh et al. requires a Taylor expansion of the likelihood of the model up to the third order at the MLE, and the likelihood of the model we consider here at the n-th step is very obviously not continuously differentiable w.r.t. u in each observed temperature t_i, $i = 1, \dots, n$. Note that the problem only grows worse as the number of observations increases.

To overcome this difficulty we follow the original idea first introduced in Sylwester (1965), and later used again in Feder (1975): we introduce a pseudo-problem for which we are able to recover the classical results and show that the differences between the estimates for the problem and the pseudo-problem are, in a sense, negligible. The pseudo-problem is obtained by deleting all the observations within intervals D_n of respective sizes d_n centred around u_0. The intervals D_n are defined as

$$D_n = \left]u_0 - \frac{d_n}{2}, u_0 + \frac{d_n}{2}\right[,$$

and their sizes d_n are chosen such that as $n \to +\infty$

$$d_n \to 0, \qquad\qquad n^{-\frac{1}{2}}(\log n) \cdot d_n^{-1} \to 0. \tag{2.10}$$

This new problem is called pseudo-problem because the value of u_0 is unknown and we therefore cannot in practise delete these observations. Note that the actual choice of the sequence $(d_n)_{n\in\mathbb{N}}$ does not influence the rest of the results in any way, as long as it satisfies to conditions (2.10). It thus does not matter at all whether one chooses (for instance) $d_n = n^{-\frac{1}{4}}$ or $d_n = \log^{-1} n$.

Let us denote n^{**} the number of observations deleted from the original problem, and $n^* = n - n^{**}$ the sample size of the pseudo-problem. Generally speaking, quantities annotated with a single asterisk * will refer to the pseudo-problem. $l^*_{1:n}(X_{1:n}|\theta)$ will thus designate the likelihood of the pseudo-problem i.e. (reindexing observations whenever necessary)

$$l^*_{1:n}(X_{1:n}|\theta) = -\frac{n^*}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n^*}\frac{1}{2\sigma^2}\left(X_i - \gamma\cdot(t_i - u)\mathbb{1}_{[t_i,+\infty[}(u)\right)^2. \tag{2.11}$$

On one hand, from an asymptotic point of view, the removal of those n^{**} observations should not have any kind of impact on the distribution theory. The intuitive idea is that deleting n^{**} observations takes away only a fraction n^{**}/n of the information which asymptotically approaches zero as will be shown below. The first condition (2.10) seems only a natural requirement if we ever hope to prove that the MLE for the problem and the pseudo-problem behave asymptotically in a similar manner (we will show they do in Theorem 2.9, see equation (2.20)).

On the other hand, assuming the MLE is consistent (we will show it is, in Theorem 2.7) and assuming that the sizes d_n are carefully chosen so that the sequence $(\widehat{u}_n)_{n\in\mathbb{N}}$ falls into the designed sequence of intervals $(D_n)_{n\in\mathbb{N}}$ (see Proposition 2.8, whose proof the second condition (2.10) is tailored for), these regions will provide open neighbourhoods of the MLE over which the likelihood of the pseudo-problem will be differentiable. The pseudo-problem can therefore be thought of as a locally regularised version of the problem (locally because we are only interested in the differentiability of the likelihood over a neighbourhood of the MLE). We should thus be able to retrieve the usual results for the pseudo-problem with a bit of work. It will be shown that this is indeed the case (see Theorem 2.3).

If the sequence $(d_n)_{n\in\mathbb{N}}$ satisfies to conditions (2.10), then as $n \to +\infty$,

$$\frac{n^{**}}{n} \to 0, \qquad\qquad \frac{n^{*}}{n} \to 1.$$

Using the uniform convergence of F_n to F over any compact subset (see Assumption A1, and its Remark 1), we indeed find via a Taylor-Lagrange approximation

$$\begin{aligned}\frac{n^{**}}{n} &= F_n\left(u_0 + \frac{d_n}{2}\right) - F_n\left(u_0 - \frac{d_n}{2}\right)\\ &= F\left(u_0 + \frac{d_n}{2}\right) - F\left(u_0 - \frac{d_n}{2}\right) + \mathrm{o}(1)\\ &= d_n \cdot f(u_n) + \mathrm{o}(1),\end{aligned}$$

where $u_n \in D_n$, so that in the end, since $u_n \to u_0$ and f is continuous and positive at u_0, we have a.s.

$$\frac{n^{**}}{n} = d_n \cdot (f(u_0) + \mathrm{o}(1)) + \mathrm{o}(1) \to 0.$$

We now recover the asymptotic normality of the posterior distribution for the pseudo problem.

Theorem 2.3. *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

$$t^* = n^{\frac{1}{2}}(\theta - \widehat{\theta}_n^*), \tag{2.12}$$

and $\widetilde{\pi}_n^*(\cdot|X_{1:n})$ the posterior density of t^* given $X_{1:n}$, then under Assumptions A1–A4 and conditions (2.10), *for any $0 \leqslant k \leqslant k_0$, as $n \to +\infty$,*

$$\int_{\mathbb{R}^3} \|t\|^k \left|\widetilde{\pi}_n^*(t|X_{1:n}) - (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}}e^{-\frac{1}{2}t'I(\theta_0)t}\right| \mathrm{d}t \xrightarrow{a.s} 0, \tag{2.13}$$

where $I(\theta)$ is defined in (2.7).

Proof of Theorem 2.3. The extensive proof, to be found in Section 2.6.1, was inspired by that of Theorem 4.2 in Ghosh et al. (2006) which deals with the case where the observations $X_1, \dots, X_n$ are independent and identically distributed and where the (univariate) log-likelihood is differentiable in a fixed small neighbourhood of θ_0. We tweaked the original proof of Ghosh et al. so that we could deal with independent but not identically distributed observations and a (multivariate) log-likelihood that is guaranteed differentiable only on a decreasing small neighbourhood of θ_0. ■

2.2.3 From the pseudo-problem to the original problem

We now give a short proof of Theorem 2.2. As we previously announced, it relies upon its counterpart for the pseudo-problem, i.e. Theorem 2.3.

Proof of Theorem 2.2. Recalling the definition of t and t^* given in (2.8) and (2.12) we observe that

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

Thus the posterior distribution of t^* and that of t, given $X_{1:n}$ are linked together via

$$\widetilde{\pi}_n(t|X_{1:n}) = \widetilde{\pi}_n^*(t - \alpha_n|X_{1:n}) \tag{2.14}$$

where

$$\alpha_n = n^{\frac{1}{2}}(\widehat{\theta}_n^* - \widehat{\theta}_n).$$

Relationship (2.14) allows us to write

$$\begin{aligned}
&\int_{\mathbb{R}^3} \|t\|^k \left|\widetilde{\pi}_n(t|X_{1:n}) - (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2}t'I(\theta_0)t}\right| \mathrm{d}t \\
&= \int_{\mathbb{R}^3} \|t\|^k \left|\widetilde{\pi}_n^*(t - \alpha_n|X_{1:n}) - (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2}t'I(\theta_0)t}\right| \mathrm{d}t \\
&= \int_{\mathbb{R}^3} \|t + \alpha_n\|^k \left|\widetilde{\pi}_n^*(t|X_{1:n}) - (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2}(t+\alpha_n)'I(\theta_0)(t+\alpha_n)}\right| \mathrm{d}t \\
&\leqslant \int_{\mathbb{R}^3} \|t + \alpha_n\|^k \left|\widetilde{\pi}_n^*(t|X_{1:n}) - (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}} e^{-\frac{1}{2}t'I(\theta_0)t}\right| \mathrm{d}t \\
&\quad + (2\pi)^{-\frac{3}{2}}|I(\theta_0)|^{\frac{1}{2}} \int_{\mathbb{R}^3} \|t + \alpha_n\|^k \left|e^{-\frac{1}{2}(t+\alpha_n)'I(\theta_0)(t+\alpha_n)} - e^{-\frac{1}{2}t'I(\theta_0)t}\right| \mathrm{d}t
\end{aligned}$$

Theorem 2.3 ensures that the first integral on the right hand side of this last inequality goes to zero in probability. It therefore suffices to show that the second integral goes to zero in probability to end the proof, i.e. that as $n \to +\infty$

$$\int_{\mathbb{R}^3} \|t + \alpha_n\|^k \left|e^{-\frac{1}{2}(t+\alpha_n)'I(\theta_0)(t+\alpha_n)} - e^{-\frac{1}{2}t'I(\theta_0)t}\right| \mathrm{d}t \xrightarrow{\mathbb{P}} 0. \tag{2.15}$$

But the proof of (2.15) is straightforward knowing that $\alpha_n \xrightarrow{\mathbb{P}} 0$ (see (2.20)) and using dominated convergence. ■

As an immediate consequence of Theorem 2.2 we want to mention the weak consistency of the Bayes estimator.

Corollary 2.4. *Let $\pi(\cdot)$ a prior distribution on θ, continuous and positive at θ_0, such that*

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

and denote

$$\widetilde{\theta}_n = \int_\Theta \theta\pi_n(\theta|X_{1:n})\,\mathrm{d}\theta,$$

the Bayes estimator of θ in the problem. Then under Assumptions A1–A4, as $n \longrightarrow +\infty$,

$$n^{\frac{1}{2}}(\widetilde{\theta}_n - \widehat{\theta}_n) \xrightarrow{\mathbb{P}} 0.$$

Proof of Corollary 2.4. By definition,

$$\widetilde{\theta}_n = \int_\Theta \theta\pi_n(\theta|X_{1:n})\,\mathrm{d}\theta$$

and this allows us to write

$$\begin{aligned} n^{\frac{1}{2}}(\widetilde{\theta}_n - \widehat{\theta}_n) &= \int_\Theta n^{\frac{1}{2}}(\theta - \widehat{\theta}_n)\pi_n(\theta|X_{1:n})\,\mathrm{d}\theta \\ &= \int_{\mathbb{R}^3} t\widetilde{\pi}_n(t|X_{1:n})\,\mathrm{d}t \xrightarrow{\mathbb{P}} 0, \end{aligned}$$

the last convergence being a direct consequence of Theorem 2.2 with $k_0 = 1$. ∎

Observe that, under conditions (2.10), the same arguments naturally apply to the pseudo-problem and lead to a strong consistency (a.s. convergence) of its associated Bayes estimator due to Theorem 2.3, thus recovering the results of Ghosh et al. (2006) for the regularised version of the problem.

2.3 Strong consistency of the MLE

In this Section we prove the strong consistency of the MLE over any compact set including the true parameter (see Theorem 2.5). It is a prerequisite for a more accurate version of the strong consistency (see Theorem 2.7) which lies at the heart of the proof of Theorem 2.3.

Theorem 2.5. *Under Assumptions A1–A4, we have a.s., as $n \longrightarrow +\infty$,*

$$\|\widehat{\theta}_n - \theta_0\| = \mathrm{o}(1).$$

Proof of Theorem 2.5. Recall that K is a compact subset of Θ, such that $\widehat{\theta}_n \in K$ for any n large enough. We denote

$$\begin{aligned} l_{1:n}(X_{1:n}|S) &= \sup_{\theta\in S} l_{1:n}(X_{1:n}|\theta), \text{ for any } S \subset K, \\ K_n(a) &= \{\theta \in \Theta,\ l_{1:n}(X_{1:n}|\theta) \geqslant \log a + l_{1:n}(X_{1:n}|K)\}, \text{ for any } a \in]0,1[. \end{aligned}$$

All we need to prove is that

$$\exists a \in]0,1[,\ \mathbb{P}\left(\lim_{n\longrightarrow+\infty} \sup_{\theta\in K_n(a)} \|\theta - \theta_0\| = 0\right) = 1. \tag{2.16}$$

since for any n large enough we have $\widehat{\theta}_n \in K_n(a)$ for any $a \in]0,1[$. We control the likelihood upon the complement of a small ball in K and prove the contrapositive of (2.16) using compacity arguments. The extensive proof of (2.16) is to be found in Section 2.6.2 . ∎

We strengthen the result of Theorem 2.5 by giving a rate of convergence for the MLE (see Theorem 2.7). This requires a rate of convergence for the image of the MLE through the regression function of the model, that we give in the Proposition 2.6 below.

Proposition 2.6. *Under Assumptions A1–A4, as $n \to +\infty$, a.s., for any open interval $I \subset [\underline{u}, \overline{u}]$,*

$$\min_{t_i \in I,\, i \leqslant n} |\mu(\widehat{\eta}_n, t_i) - \mu(\eta_0, t_i)| = \mathrm{O}\left(n^{-\frac{1}{2}} \log n\right).$$

Proof of Proposition 2.6. The proof is given in Section 2.6.2. ■

Theorem 2.7. *Under Assumptions A1–A4, we have a.s., as $n \to +\infty$,*

$$\|\widehat{\theta}_n - \theta_0\| = \mathrm{O}\left(n^{-\frac{1}{2}} \log n\right). \tag{2.17}$$

Proof of Theorem 2.7. We show that a.s. (2.17) holds for each coordinate of $\widehat{\theta}_n - \theta_0$. The calculations for the variance σ^2 are pushed back into Section 2.6.2. We now prove the result for the parameters γ and u. It is more convenient to use a reparametrisation of the model in terms of slope γ and intercept β where $\beta = -\gamma u$.

Slope γ and intercept β. Let V_1 and V_2 be two non empty open intervals of $]\underline{u}, u_0[$ such that their closures $\overline{V_1}$ and $\overline{V_2}$ do not overlap. For any $(t_1, t_2) \in V_1 \times V_2$, define $M(t_1, t_2)$ the obviously invertible matrix

$$M(t_1, t_2) = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \end{bmatrix},$$

and observe that for any $\tau = (\beta, \gamma)$,

$$M(t_1, t_2)\tau = \begin{bmatrix} \mu(\eta, t_1) \\ \mu(\eta, t_2) \end{bmatrix}.$$

Observe that by some basic linear algebra tricks we are able to write for any $(t_1, t_2) \in V_1 \times V_2$

$$\begin{aligned} \|\widehat{\tau}_n - \tau_0\|_\infty &= \|M(t_1, t_2)^{-1} M(t_1, t_2)(\widehat{\tau}_n - \tau_0)\|_\infty \\ &\leqslant |||M(t_1, t_2)^{-1}|||_\infty \cdot \|M(t_1, t_2)\widehat{\tau}_n - M(t_1, t_2)\tau_0\|_\infty \\ &\leqslant \frac{|t_2| + |t_1| + 2}{|t_2 - t_1|} \cdot \|M(t_1, t_2)\widehat{\tau}_n - M(t_1, t_2)\tau_0\|_\infty. \end{aligned}$$

Thus, using the equivalence of norms and a simple domination of the first term of the product in the inequality above, we find that there exists a constant $C \in \mathbb{R}_+^*$, such that for any $(t_1, t_2) \in V_1 \times V_2$

$$\|\widehat{\tau}_n - \tau_0\| \leqslant C \cdot \|M(t_1, t_2)\widehat{\tau}_n - M(t_1, t_2)\tau_0\|,$$

i.e.

$$\|\widehat{\tau}_n - \tau_0\| \leqslant C \cdot \left[\sum_{i=1}^{2} (\mu(\widehat{\eta}_n, t_i) - \mu(\eta_0, t_i))^2\right]^{\frac{1}{2}}. \tag{2.18}$$

Taking advantage of Proposition 2.6, we are able to exhibit two sequences of points $(t_{1,n})_{n\in\mathbb{N}}$ in V_1 and $(t_{2,n})_{n\in\mathbb{N}}$ in V_2 such that a.s., for $i = 1, 2$

$$|\mu(\widehat{\eta}_n, t_{i,n}) - \mu(\eta_0, t_{i,n})| = \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right). \tag{2.19}$$

Combining (2.18) and (2.19) together (using $t_i = t_{i,n}$ for every n), it is now trivial to see that a.s.

$$\|\widehat{\tau}_n - \tau_0\| = \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right),$$

which immediately implies the result for the γ and β components of θ.

Break-point u. Recalling that $u = -\beta\gamma^{-1}$ and thanks to the result we just proved, we find that a.s.

$$\begin{aligned}\widehat{u}_n &= -\widehat{\beta}_n\widehat{\gamma}_n^{-1} = -\left[\beta_0 + \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right)\right]\left[\gamma_0 + \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right)\right]^{-1} \\ &= -\beta_0\gamma_0^{-1} + \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right) = u_0 + \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right).\end{aligned}$$

■

2.4 Asymptotic distribution of the MLE

In this Section we derive the asymptotic distribution of the MLE for the pseudo-problem (see Proposition 2.8) and then show that the MLE of pseudo-problem and that of the problem share the same asymptotic distribution (see Theorem 2.9).

Proposition 2.8. *Under Assumptions A1–A4 and conditions* (2.10)*, as $n \to +\infty$*

$$n^{\frac{1}{2}}\left(\widehat{\theta}_n^* - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_0)^{-1}\right),$$

where the asymptotic Fisher Information Matrix $I(\cdot)$ is defined in (2.7)*.*

Proof of Theorem 2.8. The proof is divided in two steps. We first show that the likelihood of the pseudo-problem is a.s. differentiable in a neighbourhood of the MLE $\widehat{\theta}_n^*$ for N large enough. We then recover the asymptotic distribution of the MLE following the usual scheme of proof, with a Taylor expansion of the likelihood of the pseudo-problem around the true parameter. The details of these two steps are given in Section 2.6.3. ■

Theorem 2.9. *Under Assumptions A1–A4 and conditions* (2.10)*, as $n \to +\infty$,*

$$n^{\frac{1}{2}}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_0)^{-1}\right),$$

where the asymptotic Fisher Information Matrix $I(\cdot)$ is defined in (2.7)*.*

Proof of Theorem 2.9. It is a direct consequence of Proposition 2.8 as soon as we show that as $n \to +\infty$

$$\widehat{\theta}_n - \widehat{\theta}_n^* = \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right). \tag{2.20}$$

To prove (2.20), we study each coordinate separately. For γ and u, we apply Lemmas 4.12 and 4.16 found in Feder (1975) with a slight modification: the rate of convergence d_n he uses may differ from ours but it suffices to formally replace $(\log\log n)^{\frac{1}{2}}$ by $(\log n)$ all throughout his paper and the proofs he provides go through without any other change. We thus get

$$\widehat{\gamma}_n - \widehat{\gamma}_n^* = o_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right), \qquad \widehat{u}_n - \widehat{u}_n^* = o_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right). \tag{2.21}$$

It now remains to show that

$$\widehat{\sigma}_n^2 - \widehat{\sigma}_n^{2*} = o_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right). \tag{2.22}$$

To do so, we use (2.21) and the decomposition (2.61)

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \nu_i^2(\widehat{\eta}_n) + \frac{2}{n}\sum_{i=1}^{n} \nu_i(\widehat{\eta}_n)\xi_i + \frac{1}{n}\sum_{i=1}^{n} \xi_i^2,$$

where $\nu_i(\widehat{\eta}_n) = \gamma_0 \cdot (t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \widehat{\gamma}_n \cdot (t_i - \widehat{u}_n)\mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)$. The details of this are available in Section 2.6.3. ■

2.5 DISCUSSION

In this Section, we summarise the results presented in this paper. The consistency of the posterior distribution for a piecewise linear regression model is derived as well as its asymptotic normality with suitable normalisation. The proofs of these convergence results rely on the convergence of the MLE which is also proved here. In order to obtain all the asymptotic results, a regularised version of the problem at hand, called pseudo-problem, is first studied and the difference between this pseudo-problem and the (full) problem is then shown to be asymptotically negligible.

The trick of deleting observations in a diminishing neighbourhood of the true parameter, originally found in Sylwester (1965) allows the likelihood of the pseudo-problem to be differentiated at the MLE, once the MLE is shown to asymptotically belong to that neighbourhood (this requires at least a small control of the rate of convergence of the MLE). This is the key argument needed to derive the asymptotic distribution of the MLE through the usual Taylor expansion of the likelihood at the MLE. Extending the results of Ghosh et al. (2006) to a non i.i.d. setup, the asymptotic normality of the posterior distribution for the pseudo-problem is then recovered from that of the MLE, and passes on almost naturally to the (full) problem.

The asymptotic normality of the MLE and the posterior distribution are proved in this paper in a non i.i.d. setup with a non continuously differentiable likelihood. In both cases we obtain the same asymptotic results as for an i.i.d. regular model: the rate of convergence is $\sqrt{n}$ and the limiting distribution is Gaussian (see Ghosh et al., 2006; Lehmann, 2004). For the piecewise linear regression model, the exogenous variable $t_{1:n}$ does not appear in the expression of the rate of convergence as opposed to what is known for the usual linear regression model (see Lehmann, 2004): this is due to our own Assumption A1 which implies that $t'_{1:n}t_{1:n}$ is equivalent to n. Note that for a simple linear regression model, we also obtain the rate $\sqrt{n}$ under Assumption A1. In the literature, several papers already highlighted the fact that the rate of convergence and the limiting distribution (when it exists) may be different for non regular models in the sense that the likelihood is either non continuous, or non continuously differentiable, or admits singularities (see Dacunha-Castelle, 1978; Ghosal and Samanta, 1995; Ghosh et al., 1994; Ibragimov and Khasminskii, 1981). For the piecewise regression model, the likelihood is continuous but non continuously differentiable on a countable set (but the left and right derivatives exist and are finite): the rate of convergence $\sqrt{n}$ is not so surprising in our case, because this rate was already obtained for a univariate i.i.d. model the likelihood of which has the same kind of non regularity at a single point. In that case, the rate of convergence of the MLE is shown to be $\sqrt{n}$ (see Dacunha-Castelle, 1978, for instance).

2.6 Extensive proofs

2.6.1 Proofs of Section 2.2

Proof of Theorem 2.1. To prove (2.6), we proceed as announced and deal with numerator and denominator in turn.

Majoration. From Proposition 2.21 with $\rho_n = 1$, for any given $\epsilon > 0$, we can choose $\delta > 0$ small enough so that a.s. for any n large enough

$$\sup_{\theta \in B^c(\theta_0,\delta)} \frac{1}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] \leqslant -\epsilon.$$

We thus obtain a.s. for any n large enough

$$\begin{aligned} 0 &\leqslant \int_{B^c(\theta_0,\delta)} \pi(\theta) \exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] \, \mathrm{d}\theta \\ &\leqslant e^{-n\epsilon} \int_{B^c(\theta_0,\delta)} \pi(\theta) \, \mathrm{d}\theta. \end{aligned} \tag{2.23}$$

Minoration. Define $\theta_n \in \overline{B(\theta_0,\delta)}$ such that

$$\inf_{\theta \in B(\theta_0,\delta)} \frac{1}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] = \frac{1}{n}[l_{1:n}(X_{1:n}|\theta_n) - l_{1:n}(X_{1:n}|\theta_0)]$$

It is possible to define such a θ_n because $\overline{B(\theta_0,\delta)}$ is a compact subset of Θ for $\delta > 0$ small enough and $l_{1:n}(X_{1:n}|\cdot)$ is continuous as a function of θ. Let now

$$b_n(\theta) = \left(\frac{\sigma_0^2}{\sigma^2} - 1 - \log \frac{\sigma_0^2}{\sigma^2} \right) + \frac{1}{\sigma^2} \cdot \frac{1}{n} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 . \tag{2.24}$$

Recalling the definition of the log-likelihood given in (2.2) and replacing X_i by its expression given in (2.1) we find via straightforward algebra

$$\begin{aligned}
\frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] &= \log\frac{\sigma_0^2}{\sigma^2} + \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma^2}\right)\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2\right) \\
&\quad - \frac{1}{n\sigma^2}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 - \frac{2}{\sigma^2}\frac{1}{n}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta, t_i)]\xi_i \\
&= \log\frac{\sigma_0^2}{\sigma^2} + \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma^2}\right)\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \sigma_0^2 + \sigma_0^2\right) \\
&\quad - \frac{1}{n\sigma^2}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 - \frac{2}{\sigma^2}\frac{1}{n}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta, t_i)]\xi_i \\
&= \left(\log\frac{\sigma_0^2}{\sigma^2} + 1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{\sigma^2 - \sigma_0^2}{\sigma^2\sigma_0^2}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \sigma_0^2\right) \\
&\quad - \frac{1}{n\sigma^2}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 - \frac{2}{\sigma^2}\frac{1}{n}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta, t_i)]\xi_i \qquad (2.25) \\
&= -b_n(\theta) + \frac{\sigma^2 - \sigma_0^2}{\sigma^2\sigma_0^2}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \sigma_0^2\right) - \frac{2}{\sigma^2}\frac{1}{n}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta, t_i)]\xi_i. \qquad (2.26)
\end{aligned}$$

It is now easy to see that

$$\begin{aligned}
\inf_{\theta\in B(\theta_0,\delta)} \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] &= \frac{2}{n}[l_{1:n}(X_{1:n}|\theta_n) - l_{1:n}(X_{1:n}|\theta_0)] \\
&= -b_n(\theta_n) + \frac{\sigma_n^2 - \sigma_0^2}{\sigma_n^2\sigma_0^2}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \sigma_0^2\right) - \frac{2}{\sigma_n^2}\frac{1}{n}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta_n, t_i)]\xi_i \\
&= -b_n(\theta_n) + \frac{1}{\sigma_n^2}\left[\frac{\sigma_n^2 - \sigma_0^2}{\sigma_0^2}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \sigma_0^2\right) - \frac{2}{n}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta_n, t_i)]\xi_i\right] \\
&= -b_n(\theta_n) + \frac{1}{\sigma_n^2}R_n \\
&= \left(\log\frac{\sigma_0^2}{\sigma_n^2} + 1 - \frac{\sigma_0^2}{\sigma_n^2}\right) - \frac{1}{\sigma_n^2}\cdot\frac{1}{n}\sum_{i=1}^n [\mu(\eta_0, t_i) - \mu(\eta_n, t_i)]^2 + \frac{1}{\sigma_n^2}R_n
\end{aligned}$$

where $R_n \xrightarrow{a.s} 0$ because of the Law of Large Numbers and Lemma 2.16. Thanks to Lemma 2.13 we thus find that there exists $C \in \mathbb{R}_+^*$ such that

$$\begin{aligned}
\inf_{\theta\in B(\theta_0,\delta)} \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] \geqslant &\left(\log\frac{\sigma_0^2}{\sigma_n^2} + 1 - \frac{\sigma_0^2}{\sigma_n^2}\right) \\
&- \frac{1}{\sigma_n^2}\left(C\|\theta_n - \theta_0\|^2 - R_n\right)
\end{aligned}$$

We now choose $\kappa > 0$ and $\delta > 0$ small enough so that

$$\sigma_n^2\left(\log\frac{\sigma_0^2}{\sigma_n^2} + 1 - \frac{\sigma_0^2}{\sigma_n^2}\right) \geqslant -\kappa, \qquad (2.27)$$

$$-\frac{3(\kappa + C\delta^2)}{2(\sigma_0^2 - \delta)} \geqslant -\frac{1}{2}\epsilon. \qquad (2.28)$$

Thanks to (2.27) and the definition of θ_n, we can now write that

$$\begin{aligned}\inf_{\theta\in B(\theta_0,\delta)} \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] &\geqslant -\frac{1}{\sigma_n^2}\left(\kappa + C\|\theta_n - \theta_0\|^2 - R_n\right)\\ &\geqslant -\frac{1}{\sigma_n^2}\left(\kappa + C\delta^2 - R_n\right).\end{aligned}$$

Since for any n large enough

$$|R_n| \leqslant \frac{1}{2}\left(\kappa + C\delta^2\right),$$

we find via (2.28) that for any n large enough

$$\begin{aligned}\inf_{\theta\in B(\theta_0,\delta)} \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] &\geqslant -\frac{3}{2\sigma_n^2}\left(\kappa + C\delta^2\right)\\ &\geqslant -\frac{3(\kappa + C\delta^2)}{2(\sigma_0^2 - \delta)} \geqslant -\frac{1}{2}\epsilon.\end{aligned}$$

We just proved that for any $\epsilon > 0$, we have a.s. for any n large enough

$$0 \geqslant \inf_{\theta\in B(\theta_0,\delta)} \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] \geqslant -\frac{1}{2}\epsilon,$$

which immediately implies

$$\int_{B(\theta_0,\delta)} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta \geqslant e^{-\frac{1}{2}n\epsilon}\int_{B(\theta_0,\delta)} \pi(\theta)\,\mathrm{d}\theta. \tag{2.29}$$

Conclusion. Let now $\epsilon > 0$ and $\delta > 0$ small enough so that a.s. for any n large enough (2.23) and (2.29) both hold. We have a.s. for any n large enough

$$\frac{\int_{B^c(\theta_0,\delta)} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta}{\int_{B(\theta_0,\delta)} \pi(\theta)\exp[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)]\,\mathrm{d}\theta} \leqslant \frac{\int_{B^c(\theta_0,\delta)} \pi(\theta)\,\mathrm{d}\theta}{\int_{B(\theta_0,\delta)} \pi(\theta)\,\mathrm{d}\theta} e^{-\frac{1}{2}n\epsilon} \to 0,$$

which ends the proof. ■

Proof of Theorem 2.3. Because the posterior distribution of θ in the pseudo-problem, $\pi_n^*(\cdot|X_{1:n})$, can be written as

$$\pi_n^*(\theta|X_{1:n}) \propto \pi(\theta)\exp[l_{1:n}^*(X_{1:n}|\theta)],$$

the posterior density of $t^* = n^{\frac{1}{2}}(\theta - \widehat{\theta}_n^*) \in \mathbb{R}^3$ can be written as

$$\widetilde{\pi}_n^*(t|X_{1:n}) = C_n^{-1}\pi(\widehat{\theta}_n^* + n^{-\frac{1}{2}}t)\exp[l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^* + n^{-\frac{1}{2}}t) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*)]$$

where

$$C_n = \int_{\mathbb{R}^3} \pi(\widehat{\theta}_n^* + n^{-\frac{1}{2}}t)\exp[l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^* + n^{-\frac{1}{2}}t) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*)]\,\mathrm{d}t. \tag{2.30}$$

Denoting

$$\begin{aligned}g_n(t) = {}&\pi(\widehat{\theta}_n^* + n^{-\frac{1}{2}}t)\exp[l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^* + n^{*-\frac{1}{2}}t) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*)]\\ &- \pi(\theta_0)e^{-\frac{1}{2}t'I(\theta_0)t},\end{aligned} \tag{2.31}$$

to prove (2.13) it suffices to show that for any $0 \leqslant k \leqslant k_0$,

$$\int_{\mathbb{R}^3} \|t\|^k |g_n(t)| \, \mathrm{d}t \xrightarrow{a.s} 0. \tag{2.32}$$

Indeed, if (2.32) holds, $C_n \xrightarrow{a.s} \pi(\theta_0)(2\pi)^{\frac{3}{2}}|I(\theta_0)|^{-\frac{1}{2}}$ ($k = 0$) and therefore, the integral in (2.13) which is dominated by

$$\begin{aligned} C_n^{-1} &\int_{\mathbb{R}^3} \|t\|^k |g_n(t)| \, \mathrm{d}t \\ &+ \int_{\mathbb{R}^3} \|t\|^k \left| C_n^{-1}\pi(\theta_0)e^{-\frac{1}{2}t' I(\theta_0)t} - (2\pi)^{-\frac{1}{2}}|I(\theta_0)|^{\frac{1}{2}}e^{-\frac{1}{2}t' I(\theta_0)t} \right| \, \mathrm{d}t \end{aligned}$$

also goes to zero a.s.

Let $0 < \delta$ to be chosen later, and let $0 \leqslant k \leqslant k_0$. To show (2.32), we break $\mathbb{R}^3$ into two regions

$$\begin{aligned} T_1(\delta) &= B^c(0, \delta n^{\frac{1}{2}} d_n) = \{t : \|t\| \geqslant \delta n^{\frac{1}{2}} d_n\} \\ T_2(\delta) &= B(0, \delta n^{\frac{1}{2}} d_n) = \{t : \|t\| < \delta n^{\frac{1}{2}} d_n\} \end{aligned}$$

and show that for $i = 1, 2$

$$\int_{T_i(\delta)} \|t\|^k |g_n(t)| \, \mathrm{d}t \xrightarrow{a.s} 0. \tag{2.33}$$

Proof for $i = 1$. Note that $\int_{T_1(\delta)} \|t\|^k |g_n(t)|$ is dominated by

$$\begin{aligned} &\int_{T_1(\delta)} \|t\|^k \pi(\widehat{\theta}_n^* + n^{\frac{1}{2}} t) \exp[l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^* + n^{-\frac{1}{2}} t) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*)] \, \mathrm{d}t \\ &\quad + \int_{T_1(\delta)} \|t\|^k \pi(\theta_0) e^{-\frac{1}{2}t' I(\theta_0)t} \, \mathrm{d}t. \end{aligned}$$

The second integral trivially goes to zero. For the first integral, we observe that it can be rewritten as

$$n^{\frac{1}{2}} \int_{B^c(\widehat{\theta}_n^*, \delta d_n)} n^{\frac{k}{2}} \|\theta - \widehat{\theta}_n^*\|^k \pi(\theta) \exp[l_{1:n}^*(X_{1:n}|\theta) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*)] \, \mathrm{d}\theta.$$

The strong consistency of $\widehat{\theta}_n^*$ (see Theorem 2.7) implies that a.s., for any n large enough

$$\|\widehat{\theta}_n^* - \theta_0\| < \frac{1}{2}\delta d_n.$$

From this, we deduce that a.s., for any n large enough, $B^c(\widehat{\theta}_n^*, \delta d_n) \subset B^c(\theta_0, \frac{1}{2}\delta d_n)$ and thus that the first integral is dominated by

$$n^{\frac{k+1}{2}} \int_{B^c(\theta_0, \frac{1}{2}\delta d_n)} \|\theta - \widehat{\theta}_n^*\|^k \pi(\theta) \exp[l_{1:n}^*(X_{1:n}|\theta) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*)] \, \mathrm{d}\theta.$$

Recalling that $n^* \sim n$, Proposition 2.21 with $\rho_n = d_n$ implies that there a.s. exists $\epsilon > 0$ such that for any n large enough and any $\theta \in B^c(\theta_0, \frac{1}{2}\delta d_n)$ we have

$$l_{1:n}^*(X_{1:n}|\theta) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*) \leqslant -\epsilon n d_n^2.$$

It follows, using (2.10) that, a.s. for any n large enough the first integral is dominated by

$$\begin{aligned} n^{\frac{k+1}{2}} \exp(-\epsilon n d_n^2) \int_{\Theta} \|\theta - \widehat{\theta}_n^*\|^k \pi(\theta) \, \mathrm{d}t &= n^{\frac{k+1}{2}} \exp(-\epsilon n d_n^2) \cdot \mathrm{O}(1) \\ &\leqslant n^{\frac{k+1}{2}} n^{-\epsilon \log n} \cdot \mathrm{O}(1) \to 0, \end{aligned}$$

since by (2.10) we find that $n d_n^2 \geqslant (\log n)^2$ for any n large enough. Hence (2.33) holds for $i = 1$.

Proof for $i = 2$. We first recall the multivariate Taylor expansion for a function g (k+1)-times continuously differentiable within a neighbourhood of $y \in \mathbb{R}^n$. With the usual differential calculus notations

$$\mathrm{D}^\alpha g(y) \cdot h^{(\alpha)} = \sum_{1 \leqslant i_1, \ldots, i_\alpha \leqslant n} \frac{\partial^\alpha g}{\partial_{i_1} \cdots \partial_{i_\alpha}}(y) \cdot h_{i_1} \cdots h_{i_\alpha}$$

we have

$$g(x) = \sum_{\alpha=0}^{k} \frac{1}{\alpha!} \mathrm{D}^\alpha g(y) \cdot (x-y)^{(\alpha)} + R_{k+1}(x) \tag{2.34}$$

where

$$R_{k+1}(x) = \frac{1}{(k+1)!} \int_0^1 (1-s)^k \mathrm{D}^{k+1} g(y + s(x-y)) \cdot (x-y)^{(k+1)} \, \mathrm{d}s. \tag{2.35}$$

Before expanding the log-likelihood over $T_2(\delta)$ in a such a way, we first have to make sure it is differentiable over the correct domain. Indeed, the strong consistency of $\widehat{\theta}_n^*$ (see Theorem 2.7) implies that a.s., whatever $\delta_0 > 0$, for n large enough,

$$\|\widehat{\theta}_n^* - \theta_0\| < \delta_0 d_n.$$

For δ chosen small enough, since $t \in T_2(\delta)$ implies

$$\|\theta - \widehat{\theta}_n^*\| < \delta d_n$$

it follows from the triangle inequality that a.s. for n large enough,

$$\|\theta - \theta_0\| < (\delta_0 + \delta) d_n < d_n.$$

A.s. for any n large enough, $t \in T_2(\delta)$ hence implies $\theta \in B(\theta_0, (\delta + \delta_0) d_n)$. We choose δ_0 and δ small enough so that $\delta + \delta_0 < 1$. This way, $\theta \mapsto l_{1:n}^*(X_{1:n}|\theta)$ is guaranteed to be infinitely continuously differentiable over $B(\theta_0, (\delta + \delta_0) d_n) \subset B(\theta_0, d_n)$.

Now expanding the log-likelihood in a Taylor series for any n large enough, and taking advantage of the fact that $l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*) = 0$, we define $B_{1:n}^*(\cdot)$ the symmetric matrix defined for $u \in D_n$ by

$$B_{1:n}^*(\theta) = - \begin{bmatrix} \dfrac{\partial^2 l_{1:n}^*(X_{1:n}|\theta)}{\partial\gamma\partial\gamma} & \dfrac{\partial^2 l_{1:n}^*(X_{1:n}|\theta)}{\partial\gamma\partial u} & \dfrac{\partial^2 l_{1:n}^*(X_{1:n}|\theta)}{\partial\gamma\partial\sigma^2} \\ & \dfrac{\partial^2 l_{1:n}^*(X_{1:n}|\theta)}{\partial u\partial u} & \dfrac{\partial^2 l_{1:n}^*(X_{1:n}|\theta)}{\partial u\partial\sigma^2} \\ & & \dfrac{\partial^2 l_{1:n}^*(X_{1:n}|\theta)}{\partial\sigma^2\partial\sigma^2} \end{bmatrix}. \tag{2.36}$$

and write that

$$\begin{aligned} l_{1:n}^*(X_{1:n}|\theta) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*) = & -\frac{1}{2}(\theta - \widehat{\theta}_n^*)' \left(B_{1:n}^*(\widehat{\theta}_n^*)\right)(\theta - \widehat{\theta}_n^*) \\ & + R_{3,n}(\theta) \end{aligned} \tag{2.37}$$

where

$$R_{3,n}(\theta) = \frac{1}{3!} \int_0^1 (1-s)^2 \mathrm{D}^3 l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^* + s(\theta - \widehat{\theta}_n^*)) \cdot (\theta - \widehat{\theta}_n^*)^{(3)} \, \mathrm{d}s. \tag{2.38}$$

Lemma 2.22 allows us to write that a.s. there exists a constant $C \in R_+^*$ such that for any n large enough, for any $t \in T_2(\delta)$

$$l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^* + n^{-\frac{1}{2}}t) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*) = -\frac{1}{2}t'\left(n^{-1}B_{1:n}^*(\widehat{\theta}_n^*)\right)t + S_n(t) \tag{2.39}$$

where

$$|S_n(t)| \leqslant Cn^{-\frac{1}{2}} \cdot \|t\|^3. \tag{2.40}$$

From (2.40), we obtain that for any $t \in T_2(\delta)$, $S_n(t) \xrightarrow{a.s} 0$. Because of Lemma 2.20, we have $n^{-1}B_{1:n}^*(\widehat{\theta}_n^*) \xrightarrow{a.s} I(\theta_0)$, and it follows immediately that for any $t \in T_2(\delta)$,

$$g_n(t) \xrightarrow{a.s} 0,$$

and thus that

$$\|t\|^k g_n(t) \xrightarrow{a.s} 0.$$

From (2.40) we also obtain

$$|S_n(t)| \leqslant C\delta d_n \|t\|^2.$$

Lemma 2.20, combined with (2.10), (2.39) and the positivity of $I(\theta_0)$, ensures that a.s. for any n large enough

$$|S_n(t)| \leqslant \frac{1}{4}t'\left(n^{-1}B_{1:n}^*(\widehat{\theta}_n^*)\right)t,$$

so that from (2.39), a.s. for any n large enough

$$\exp[l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^* + n^{-\frac{1}{2}}t) - l_{1:n}^*(X_{1:n}|\widehat{\theta}_n^*)] \leqslant e^{-\frac{1}{4}t'\left(n^{-1}B_{1:n}^*(\widehat{\theta}_n^*)\right)t} \leqslant e^{-\frac{1}{8}t'I(\theta_0)t}. \tag{2.41}$$

Therefore, for n large enough, $\|t\|^k|g_n(t)|$ is dominated by an integrable function on the set $T_2(\delta)$ and (2.33) holds for $i = 2$ which completes the proof. ■

2.6.2 Proofs of Section 2.3

Proof of Theorem 2.5. From (2.26), it is easy to see that

$$\begin{aligned}\frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|K)] &\leqslant \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] \\ &\leqslant -b_n(\theta) + \frac{\sigma^2 - \sigma_0^2}{\sigma^2\sigma_0^2}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \sigma_0^2\right) \\ &\quad - \frac{2}{\sigma^2}\frac{1}{n}\sum_{i=1}^n[\mu(\eta_0, t_i) - \mu(\eta, t_i)]\xi_i.\end{aligned}$$

For any $\theta' \in \Theta$ and $r > 0$, let $B(\theta', r) = \{\theta, ;\ \|\theta' - \theta\|_1 < r\}$. It is now obvious that

$$\begin{aligned}&\frac{2}{n}[l_{1:n}(X_{1:n}|B(\theta', r)) - l_{1:n}(X_{1:n}|K)] \\ &\leqslant \sup_{\theta\in B(\theta',r)}\{-b_n(\theta)\} + \sup_{\theta\in B(\theta',r)}\left|\frac{\sigma^2 - \sigma_0^2}{\sigma^2\sigma_0^2}\right| \cdot \left|\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \sigma_0^2\right| \\ &\quad + \sup_{\theta\in B(\theta',r)}\left\{\frac{2}{\sigma^2}\right\} \cdot \sup_{\theta\in B(\theta',r)}\left\{\left|\frac{1}{n}\sum_{i=1}^n[\mu(\eta_0, t_i) - \mu(\eta, t_i)]\xi_i\right|\right\}.\end{aligned} \tag{2.42}$$

Lemma 2.16 now ensures that

$$\sup_{\theta\in B(\theta',r)}\left|\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_0,t_i)-\mu(\eta,t_i)]\xi_i\right| \xrightarrow{a.s} 0,$$

and σ^2 being bounded away from 0 ensures the boundedness of $\sup_{\theta\in B(\theta',r)}\left\{\frac{2}{\sigma^2}\right\}$ which implies

$$\sup_{\theta\in B(\theta',r)}\left\{\frac{2}{\sigma^2}\right\}\cdot\sup_{\theta\in B(\theta',r)}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_0,t_i)-\mu(\eta,t_i)]\xi_i\right|\right\} \xrightarrow{a.s} 0.$$

Since σ^2 is bounded away from 0, taking advantage of the Strong Law of Large Numbers, we also obtain

$$\sup_{\theta\in B(\theta',r)}\left|\frac{\sigma^2-\sigma_0^2}{\sigma^2\sigma_0^2}\right|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i^2-\sigma_0^2\right| \xrightarrow{a.s} 0.$$

We may thus rewrite (2.42) as

$$\frac{2}{n}[l_{1:n}(X_{1:n}|B(\theta',r))-l_{1:n}(X_{1:n}|K)] \leqslant \sup_{\theta\in B(\theta',r)}\{-b_n(\theta)\}+R_n, \tag{2.43}$$

where $R_n \xrightarrow{a.s} 0$.

Assume now that $\theta'\neq\theta_0$, then we have

$$\sup_{\theta\in B(\theta',r)}|b_n(\theta)-b(\theta')| \leqslant \sup_{\theta\in B(\theta',r)}|b_n(\theta)-b_n(\theta')|+|b_n(\theta')-b(\theta')|. \tag{2.44}$$

Lemma 2.15 (see (2.79)) ensures the existence of a r small enough, say $r=r(\theta')$, such that

$$\sup_{\theta\in B(\theta',r(\theta'))}|b_n(\theta)-b_n(\theta')| \leqslant \frac{1}{4}b(\theta'), \tag{2.45}$$

uniformly in n. For n large enough, that same Lemma 2.15 (see (2.80)) also guarantees that

$$|b_n(\theta')-b(\theta')| \leqslant \frac{1}{4}b(\theta'). \tag{2.46}$$

Adding inequalities (2.45) and (2.46) together and combining the result with (2.44), we deduce that for any n large enough

$$\sup_{\theta\in B(\theta',r(\theta'))}|b_n(\theta)-b(\theta')| \leqslant \frac{1}{2}b(\theta'),$$

i.e.

$$\sup_{\theta\in B(\theta',r(\theta'))}\{-b_n(\theta)\} \leqslant -\frac{1}{2}b(\theta'),$$

which finally gives together with (2.43)

$$\forall\theta'\neq\theta_0,\ \mathbb{P}\left(\limsup_{n\longrightarrow+\infty}\frac{1}{n}[l_{1:n}(X_{1:n}|B(\theta',r(\theta')))-l_{1:n}(X_{1:n}|K)] \leqslant -\frac{1}{4}b(\theta')\right)=1. \tag{2.47}$$

Since Lemma 2.15 ensures that $b(\theta') > 0$ for any $\theta' \neq \theta_0$, the previous statement implies

$$\forall \theta' \neq \theta_0,\ \mathbb{P}\left(\exists n(\theta') \in \mathbb{N},\ \forall n > n(\theta'),\ l_{1:n}(X_{1:n}|B(\theta', r(\theta'))) - l_{1:n}(X_{1:n}|K) < -1\right) = 1. \qquad (2.48)$$

For a given $\delta > 0$, let us now define $K(\delta) = K \setminus B(\theta_0, \delta)$. $K(\delta)$ is obviously a compact set since K itself is a compact set. By compacity, from the covering

$$\bigcup_{\theta' \in K(\delta)} B(\theta', r(\theta')) \supset K(\delta),$$

there exists a finite subcovering, i.e.

$$\exists m(\delta) \in \mathbb{N},\ \bigcup_{j=1}^{m(\delta)} B(\theta'_j, r(\theta'_j)) \supset K(\delta).$$

In particular, (2.48) holds for $\theta' = \theta'_j, j = 1, \ldots, m(\delta)$. Let us define

$$n_0(\delta) = \max_{j=1,\ldots,m(\delta)} n(\theta'_j).$$

We may now write

$$\begin{aligned}&\forall \delta > 0,\ \exists n_0(\delta) \in \mathbb{N},\ \exists m(\delta) \in \mathbb{N},\ \forall j = 1, \ldots, m(\delta),\\ &\qquad \mathbb{P}\Big(\forall n > n_0(\delta),\ l_{1:n}(X_{1:n}|B(\theta'_j, r(\theta'_j))) - l_{1:n}(X_{1:n}|K) < -1\Big) = 1,\end{aligned}$$

which we turn into

$$\begin{aligned}&\forall \delta > 0,\ \exists n_0(\delta) \in \mathbb{N},\ \exists m(\delta) \in \mathbb{N},\\ &\qquad \mathbb{P}\Big(\forall n > n_0(\delta),\ \forall j = 1, \ldots, m(\delta),\ l_{1:n}(X_{1:n}|B(\theta'_j, r(\theta'_j))) - l_{1:n}(X_{1:n}|K) < -1\Big) = 1,\end{aligned}$$

thanks to the finiteness of $m(\delta)$, and finally into

$$\forall \delta > 0,\ \exists n_0(\delta) \in \mathbb{N},\ \mathbb{P}\left(\forall n > n_0(\delta),\ l_{1:n}(X_{1:n}|K(\delta)) - l_{1:n}(X_{1:n}|K) < -1\right) = 1,$$

because of the covering

$$\bigcup_{j=1}^{m(\delta)} B(\theta'_j, r(\theta'_j)) \supset K(\delta).$$

Let us now sum up what we have obtained so far. We proved that

$$\forall \delta > 0,\ \exists n_0(\delta) \in \mathbb{N},\ \mathbb{P}\Big(\text{if } \forall n > n_0(\delta),\ l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|K) \geqslant \log e^{-1},\ \text{then } \theta \notin K(\delta)\Big) = 1,$$

i.e.

$$\exists a = e^{-1} \in]0, 1[,\ \forall \delta > 0,\ \exists n_0(\delta) \in \mathbb{N},\ \mathbb{P}\left(\text{if } \forall n > n_0(\delta),\ \theta \in K_n(a),\ \text{then } \|\theta - \theta_0\|_1 < \delta\right) = 1,$$

that is to say

$$\exists a \in]0, 1[,\ \mathbb{P}\left(\lim_{n \longrightarrow +\infty} \sup_{\theta \in K_n(a)} \|\theta - \theta_0\|_1 = 0\right) = 1.$$

■

Proof of Proposition 2.6. In this proof $\| \cdot \|$ will refer to the usual Euclidean norm. Reindexing whenever necessary, we also assume that the observations t_i are ordered, and we denote

$$t = (t_1, \ldots, t_n), \qquad X = (X_1, \ldots, X_n), \qquad \mu_0 = (\mu(\eta_0, t_1), \ldots, \mu(\eta_0, t_n)),$$

$$N_{0,n} = \sup_{i \leqslant n}\{i,\ t_i < u_0\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[t_i, +\infty[}(u_0), \qquad N_n = \sup_{i \leqslant n}\{i,\ t_i < \widehat{u}_n\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[t_i, +\infty[}(\widehat{u}_n),$$

$$\zeta = \begin{cases} (0, \ldots, 0, \beta_0 + \gamma_0 t_{N_n+1}, \ldots, \beta_0 + \gamma_0 t_{N_{0,n}}, 0, \ldots, 0), & \text{if } N_n < N_{0,n} \\ (0, \ldots, 0), & \text{if } N_n = N_{0,n} \\ (0, \ldots, 0, \beta_0 + \gamma_0 t_{N_{0,n}+1}, \ldots, \beta_0 + \gamma_0 t_{N_n}, 0, \ldots, 0), & \text{if } N_n > N_{0,n} \end{cases},$$

Let $\mathcal{G}$ be the linear space spanned by the 2 linearly independent n-vectors

$$v_1 = (1, \ldots, 1, 0, \ldots, 0) \qquad v_2 = (t_1, \ldots, t_{N_n}, 0, \ldots, 0)$$

(both of which have their last $n - N_n$ coordinates valued to zero), and denote Q the orthogonal projection onto $\mathcal{G}$.

Let $\mathcal{G}^+$ denote the linear space spanned by v_1, v_2 and μ_0 and denote Q^+ the orthogonal projection onto $\mathcal{G}^+$. Observe that $\mathcal{G}^+$ is also spanned by v_1, v_2 and ζ.

Finally, denote μ^* the orthogonal projection of X onto $\mathcal{G}^+$ and $\widehat{\mu}$ the closest point to X in $\mathcal{G}^+$ satisfying the continuity assumption of the model, i.e.

$$\mu^* = Q^+ X, \qquad \widehat{\mu} = (\mu(\widehat{\eta}_n, t_1), \ldots, \mu(\widehat{\eta}_n, t_n)).$$

We have

$$\begin{aligned} \|X - \mu^*\|^2 + \|\mu^* - \widehat{\mu}||^2 = \|X - \widehat{\mu}\|^2 &\leqslant \|X - \mu_0\|^2, \\ \|X - \mu_0\|^2 - \|\mu^* - \mu_0\|^2 + \|\mu^* - \widehat{\mu}||^2 &\leqslant \|X - \mu_0\|^2, \\ \|\mu^* - \mu_0\|^2 - 2\langle \mu^* - \mu_0, \widehat{\mu} - \mu_0\rangle + \|\widehat{\mu} - \mu_0\|^2 &\leqslant \|\mu^* - \mu_0\|^2. \end{aligned}$$

Thus

$$\|\widehat{\mu} - \mu_0\|^2 \leqslant 2\langle \mu^* - \mu_0, \widehat{\mu} - \mu_0\rangle \leqslant 2\|\mu^* - \mu_0\| \cdot \|\widehat{\mu} - \mu_0\|,$$

which leads to

$$\|\widehat{\mu} - \mu_0\| \leqslant 2\|\mu^* - \mu_0\| \leqslant 2\|Q^+\zeta\|.$$

Our aim is to show that a.s.

$$\|Q^+\zeta\| = \mathrm{O}(\log n). \tag{2.49}$$

If (2.49) held, then we would have a.s. $\|\widehat{\mu} - \mu_0\| = \mathrm{O}(\log n)$ i.e. a.s.

$$\sum_{i=1}^{n} (\mu(\widehat{\eta}_n, t_i - \mu(\eta_0, t_i))^2 = \mathrm{O}\left(\log^2 n\right).$$

Hence, a.s. for any open interval $I \subset [\underline{u}, \overline{u}]$ we would have

$$\sum_{i=1}^{n} (\mu(\widehat{\eta}_n, t_i - \mu(\eta_0, t_i))^2 \, \mathbb{1}_I(t_i) = \mathrm{O}\left(\log^2 n\right).$$

This would immediately imply the desired result, i.e. that a.s.

$$\min_{t_i \in I,\, i \leqslant n} |\mu(\widehat{\eta}_n, t_i) - \mu(\eta_0, t_i)| = \mathrm{O}\left(n^{-\frac{1}{2}} \log n\right),$$

since a.s.

$$\mathrm{O}\left(\log^2 n\right) = \sum_{i=1}^{n} (\mu(\widehat{\eta}_n, t_i - \mu(\eta_0, t_i))^2 \, \mathbb{1}_I(t_i) \geqslant n \cdot \min_{t_i \in I,\, i \leqslant n} |\mu(\widehat{\eta}_n, t_i) - \mu(\eta_0, t_i)|^2 \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_I(t_i),$$

where (see Assumption A1)

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_I(t_i) = \int_I \mathrm{d}F_n(t) \to \int_I \mathrm{d}F(t) = \int_I f(t)\, \mathrm{d}t > 0.$$

Let us now prove that (2.49) indeed holds. We consider the two following mutually exclusive situations.

Situation A: $\zeta = (0, \ldots, 0)$. In this situation

$$\|Q^+ \xi\| = \|Q\xi\|, \tag{2.50}$$

and Cochran's theorem guarantees that $\|Q\xi\|^2 \sim \chi^2(2)$ for $n \geqslant 2$. Hence, via Corollary 2.17, a.s.

$$\|Q\xi\| = \mathrm{O}\left(\log n\right), \tag{2.51}$$

and (2.49) follows from (2.50) and (2.51).

Situation B: $\zeta \neq (0, \ldots, 0)$. Since

$$\frac{|\langle \zeta, \xi \rangle|}{\|\zeta\|} \sim \mathcal{N}(0, \sigma_0^2),$$

we also have, via Lemma 2.17, a.s.

$$\frac{|\langle \zeta, \xi \rangle|}{\|\zeta\|} = \mathrm{O}\left(\log n\right). \tag{2.52}$$

Notice that (2.49) follows from (2.51) and (2.52) if we manage to show that a.s.

$$\|Q^+ \xi\| \leqslant \mathrm{O}(1) \cdot \left(\|Q\xi\| + \frac{|\langle \zeta, \xi \rangle|}{\|\zeta\|}\right). \tag{2.53}$$

It thus now suffices to prove that a.s., for any $g \in \mathcal{G}$

$$|\langle \zeta, g \rangle| = \|\zeta\| \, \|g\| \cdot \mathrm{o}(1), \tag{2.54}$$

where the $\mathrm{o}(1)$ mentioned in (2.54) is uniform in g over $\mathcal{G}$ (i.e. a.s. ζ is asymptotically uniformly orthogonal to $\mathcal{G}$), for (2.53) is a direct consequence of (2.54) and Lemma 2.10 whose proof is found in Feder (1975).

Lemma 2.10. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two linear subspaces of an inner product space $\mathcal{E}$. If there exists $\alpha < 1$ such that*

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ |\langle x, y\rangle| \leqslant \alpha \|x\|\ \|y\|,$$

then

$$\|x + y\| \leqslant (1-\alpha)^{-1}(\|x^*\| + \|y^*\|),$$

where x^* (resp. y^*) is the orthogonal projection of $x + y$ onto $\mathcal{X}$ (resp. $\mathcal{Y}$).

Observe that, as a consequence of Assumption A1 and Theorem 2.11, the three following convergences are uniform in u over $[\underline{u},\ \overline{u}]$ for $k = 0, 1, 2$,

$$\frac{1}{n}\sum_{i=1}^{n} t_i^k \mathbb{1}_{[t_i,+\infty[}(u) = \int_{\underline{u}}^{u} t^k \,\mathrm{d}F_n(t) \to \int_{\underline{u}}^{u} t^k \,\mathrm{d}F(t) = \int_{\underline{u}}^{u} t^k f(t)\,\mathrm{d}t. \tag{2.55}$$

We have a.s., for any $g(\phi) = (\cos\phi)v_1 + (\sin\phi)v_2 \in \mathcal{G}$, with $\phi \in [0,\ 2\pi]$

$$\begin{aligned}
|\langle \zeta, g(\phi)\rangle| &= \left| \sum_{i=1}^{N_n} (\beta_0 + \gamma_0 t_i)(\cos\phi + t_i \sin\phi) - \sum_{i=1}^{N_{0,n}} (\beta_0 + \gamma_0 t_i)(\cos\phi + t_i \sin\phi) \right| \\
&\leqslant (\max(|\underline{u}|, |\overline{u}|) + 1) \cdot \left| \sum_{i=1}^{N_n} |\beta_0 + \gamma_0 t_i| - \sum_{i=1}^{N_{0,n}} |\beta_0 + \gamma_0 t_i| \right| \\
&\leqslant (\max(|\underline{u}|, |\overline{u}|) + 1) \cdot \|\zeta\|_1 \\
&\leqslant (\max(|\underline{u}|, |\overline{u}|) + 1) \cdot \|\zeta\| \cdot n^{\frac{1}{2}} |N_n - N_{0,n}|^{\frac{1}{2}} \\
&\leqslant (\max(|\underline{u}|, |\overline{u}|) + 1) \cdot \|\zeta\| \cdot n^{\frac{1}{2}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[t_i,+\infty[}(u_0) \right|^{\frac{1}{2}},
\end{aligned}$$

i.e. we have a.s. for any $\phi \in [0,\ 2\pi]$

$$|\langle \zeta, g(\phi)\rangle| = n^{\frac{1}{2}} \|\zeta\| \cdot \mathrm{o}(1), \tag{2.56}$$

thanks to the strong consistency $\widehat{u}_n \xrightarrow{a.s} u_0$ (see Theorem 2.5) and the uniform convergence mentioned in (2.55) with ($k = 0$). Observe that the $\mathrm{o}(1)$ mentioned in (2.56) is uniform in ϕ over $[0,\ 2\pi]$. We also have a.s. for any $\phi \in [0,\ 2\pi]$

$$\begin{aligned}
\frac{1}{n}\|g(\phi)\|^2 &= \frac{1}{n}\sum_{i=1}^{n} (\cos\phi + t_i \sin\phi)^2 \mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)\cos^2\phi + 2\frac{1}{n}\sum_{i=1}^{n} t_i \mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)\cos\phi\sin\phi \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} t_i^2 \mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)\sin^2\phi \\
&\xrightarrow{a.s} \cos^2\phi \int_{\underline{u}}^{u_0} f(t)\,\mathrm{d}t + \cos\phi\sin\phi \int_{\underline{u}}^{u_0} 2tf(t)\,\mathrm{d}t + \sin^2\phi \int_{\underline{u}}^{u_0} t^2 f(t)\,\mathrm{d}t,
\end{aligned}$$

once again making use of the strong consistency $\widehat{u}_n \xrightarrow{a.s} u_0$ (see Theorem 2.5) and taking advantage of all three uniform convergences mentioned in (2.55). We thus obviously have a.s., uniformly in ϕ over $[0,\ 2\pi]$

$$\frac{1}{n}\|g(\phi)\|^2 \to \int_{\underline{u}}^{u_0} (\cos\phi + t\sin\phi)^2 f(t)\,\mathrm{d}t. \tag{2.57}$$

The limit in (2.57) is a positive and continuous function of ϕ, and is hence bounded, i.e. there exists $m > 0$ such that we have a.s.

$$\frac{1}{n}\|g(\phi)\|^2 \geqslant m + \mathrm{o}(1), \tag{2.58}$$

i.e.

$$\frac{1}{\|g(\phi)\|} = \mathrm{O}(n^{-\frac{1}{2}}), \tag{2.59}$$

where the $\mathrm{o}(1)$ mentioned in (2.58) and the $\mathrm{O}(n^{-\frac{1}{2}})$ mentioned in (2.59) are uniform in ϕ over $[0,\, 2\pi]$.

Combining (2.56) and (2.59) together, we have a.s. for any $\phi \in [0,\, 2\pi]$

$$|\langle \zeta, g(\phi)\rangle| = \|\zeta\|\ \|g(\phi)\| \cdot \mathrm{o}(1), \tag{2.60}$$

where the $\mathrm{o}(1)$ mentioned in (2.60) is uniform in ϕ over $[0,\, 2\pi]$.

Hence, we have a.s, for any $r \in \mathbb{R}_+^*$, and any $\phi \in [0,\, 2\pi]$, now denoting $g(\phi) = (r\cos\phi)v_1 + (r\sin\phi)v_2$ and applying (2.60) to $r^{-1}g(\phi)$

$$|\langle \zeta, g(\phi)\rangle| = r\left|\left\langle \zeta, r^{-1}g(\phi)\right\rangle\right| = r\cdot\|\zeta\|\ \|r^{-1}g(\phi)\|\cdot\mathrm{o}(1) = \|\zeta\|\ \|g(\phi)\|\cdot\mathrm{o}(1),$$

where the $\mathrm{o}(1)$ mentioned is uniform in ϕ over $[0,\, 2\pi]$ and does not depend on r.

We immediately deduce that a.s. (2.54) holds i.e. a.s. ζ is asymptotically uniformly orthogonal to $\mathcal{G}$, which completes the proof. ■

Proof of Theorem 2.7. We now prove that

$$\|\widehat{\sigma}_n^2 - \sigma_0^2\| = \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right).$$

Variance of noise σ^2. Observe that

$$\begin{aligned}
\widehat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[X_i - \widehat{\gamma}_n\left(t_i - \widehat{u}_n\right)\mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)\right]^2\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\gamma_0\cdot(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \widehat{\gamma}_n\cdot(t_i - \widehat{u}_n)\mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n) + \xi_i\right]^2\\
&= \frac{1}{n}\sum_{i=1}^{n}\nu_i^2(\widehat{\eta}_n) + \frac{2}{n}\sum_{i=1}^{n}\nu_i(\widehat{\eta}_n)\xi_i + \frac{1}{n}\sum_{i=1}^{n}\xi_i^2,
\end{aligned} \tag{2.61}$$

where we denote for $i = 1, \ldots, n$,

$$\nu_i(\eta) = \gamma_0\cdot(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \gamma\cdot(t_i - u)\mathbb{1}_{[t_i,+\infty[}(u). \tag{2.62}$$

We have

$$\begin{aligned}
\sup_{i\in\mathbb{N}}|\nu_i(\widehat{\eta}_n)| &= \sup_{i\in\mathbb{N}}\left|\gamma_0\cdot(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \widehat{\gamma}_n\cdot(t_i - \widehat{u}_n)\mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)\right|\\
&\leqslant |\gamma_0 - \widehat{\gamma}_n|\cdot\sup_{i\in\mathbb{N}}\left|(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\right|\\
&\quad + |\widehat{\gamma}_n|\cdot\sup_{i\in\mathbb{N}}\left|(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - (t_i - \widehat{u}_n)\mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)\right|\\
&= \mathrm{O}\left(\gamma_0 - \widehat{\gamma}_n\right) + |\widehat{\gamma}_n|\,\mathrm{O}\left(u_0 - \widehat{u}_n\right),
\end{aligned} \tag{2.63}$$

using straightforward dominations and Lemma 2.12, so that in the end, thanks to the previous results we have a.s.

$$\sup_{i\in\mathbb{N}} |\nu_i(\widehat{\eta}_n)| = \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right). \tag{2.64}$$

It is thus easy to see that a.s.

$$\frac{1}{n}\sum_{i=1}^{n}\nu_i^2(\widehat{\eta}_n) = \mathrm{O}\left(n^{-1}\log^2 n\right) = \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right), \tag{2.65}$$

and also that, via Corollary 2.17, a.s.

$$\frac{2}{n}\sum_{i=1}^{n}\nu_i(\widehat{\eta}_n)\xi_i = \frac{2}{n}\left(\sum_{i=1}^{n}\nu_i^2(\widehat{\eta}_n)\right)^{\frac{1}{2}}\cdot \mathrm{O}(\log n) = \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right). \tag{2.66}$$

From the Law of the Iterated Logarithm (see Breiman, 1992, Chapter 13, page 291) we have a.s.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\xi_i^2-\sigma_0^2\right) = \mathrm{O}\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right) = \mathrm{O}\left(n^{-\frac{1}{2}}\log n\right) \tag{2.67}$$

and the desired result follows from (2.65), (2.66) and (2.67) put together into (2.61). ■

2.6.3 Proofs of Section 2.4

Proof of Proposition 2.8. We proceed as announced.

Step 1. We first prove that a.s.

$$\exists N\in\mathbb{N},\ \forall n>N,\ \widehat{u}_n^*\in D_n.$$

Let us notice that anything proven for the problem remains valid for the pseudo-problem. Because $n^*\sim n$, we have a.s., thanks to Theorem 2.7 and conditions (2.10), as $n\to+\infty$

$$n^{\frac{1}{2}}(\log^{-1} n)\cdot(\widehat{u}_n^*-u_0) = \mathrm{O}(1),$$
$$n^{\frac{1}{2}}(\log^{-1} n)\cdot d_n \to +\infty,$$

and thus deduce from the ratio of these two quantities that

$$\frac{\widehat{u}_n^*-u_0}{d_n}\xrightarrow{a.s}0,$$

and this directly implies the desired result.

Step 2. Let $A_{1:n}^*(\cdot)$ be the column vector defined for $u\in D_n$ by

$$A_{1:n}^*(\theta) = \left(\left.\frac{\partial l_{1:n}^*(X_{1:n}|\theta)}{\partial\gamma}\right|_{\theta}, \left.\frac{\partial l_{1:n}^*(X_{1:n}|\theta)}{\partial u}\right|_{\theta}, \left.\frac{\partial l_{1:n}^*(X_{1:n}|\theta)}{\partial\sigma^2}\right|_{\theta}\right). \tag{2.68}$$

Step 1 allows us to expand a.s. $A_{1:n}^*(\widehat{\theta}_n^*)$ around θ_0 using a Taylor-Lagrange approximation

$$0 = A_{1:n}^*(\widehat{\theta}_n^*) = A_{1:n}^*(\theta_0) - B_{1:n}^*(\widetilde{\theta}_n)\left(\widehat{\theta}_n^*-\theta_0\right),$$

where $\widetilde{\theta}_n$ is a point between $\widehat{\theta}^*_n$ and θ_0 (see (2.36) for the definitions of $B^*_{1:n}$), and rewrite it as a.s.

$$\frac{1}{n^*} B^*_{1:n}(\widetilde{\theta}_n) \cdot n^{*\frac{1}{2}} \left(\widehat{\theta}^*_n - \theta_0\right) = n^{*-\frac{1}{2}} A^*_{1:n}(\theta_0).$$

Since $\widehat{\theta}^*_n \to \theta_0$, we also have $\widetilde{\theta}_n \to \theta_0$ and using both Lemmas 2.19 and 2.20 we immediately find that as $n \to +\infty$

$$I(\theta_0) \cdot n^{*\frac{1}{2}} \left(\widehat{\theta}^*_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_0)\right),$$

which means, remembering both that $n^* \sim n$ and that $I(\theta_0)$ is positive definite and thus invertible that as $n \to +\infty$

$$n^{\frac{1}{2}} \left(\widehat{\theta}^*_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_0)^{-1}\right).$$

■

Proof of Theorem 2.9. To finish the proof, we need to show (2.22) i.e. that

$$\widehat{\sigma}^2_n - \widehat{\sigma}^{2*}_n = \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

We use the decomposition (2.61)

$$\widehat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^n \nu_i^2(\widehat{\eta}_n) + \frac{2}{n}\sum_{i=1}^n \nu_i(\widehat{\eta}_n)\xi_i + \frac{1}{n}\sum_{i=1}^n \xi_i^2,$$

where $\nu_i(\widehat{\eta}_n) = \gamma_0 \cdot (t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \widehat{\gamma}_n \cdot (t_i - \widehat{u}_n)\mathbb{1}_{[t_i,+\infty[}(\widehat{u}_n)$.

Having proved in Proposition 2.8 that

$$\widehat{\gamma}^*_n - \gamma_0 = \mathrm{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right), \qquad \widehat{u}^*_n - u_0 = \mathrm{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$$

we add these relationships to those from (2.21) and find that

$$\widehat{\gamma}_n - \gamma_0 = \mathrm{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right), \qquad \widehat{u}_n - u_0 = \mathrm{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right). \tag{2.69}$$

We now use (2.69) together with (2.63), we are able to write

$$\sup_{i\in\mathbb{N}} |\nu_i(\widehat{\eta}_n)| = \mathrm{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right). \tag{2.70}$$

It is hence easy to see that

$$\frac{1}{n}\sum_{i=1}^n \nu_i^2(\widehat{\eta}_n) = \mathrm{O}_{\mathbb{P}}\left(n^{-1}\right) = \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),$$

and also that

$$\frac{2}{n}\sum_{i=1}^n \nu_i(\widehat{\eta}_n)\xi_i = \frac{2}{n}\left(\sum_{i=1}^n \nu_i^2(\widehat{\eta}_n)\right)^{\frac{1}{2}} \cdot \mathrm{O}_{\mathbb{P}}(1) = \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),$$

which once both substituted into (2.61) yield

$$\widehat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^n \xi_i^2 + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

What was done above with the problem and $\widehat{\sigma}_n^2$ can be done with the pseudo-problem and $\widehat{\sigma}_n^{2*}$ without any kind of modification so that

$$\widehat{\sigma}_n^{2*} = \frac{1}{n^*} \sum_{i=1}^{n^*} \xi_i^2 + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

We observe that

$$\begin{aligned}
\widehat{\sigma}_n^2 - \widehat{\sigma}_n^{2*} &= \frac{1}{n}\sum_{i=1}^{n} \xi_i^2 - \frac{1}{n^*}\sum_{i=1}^{n^*} \xi_i^2 + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \\
&= \left[\frac{1}{n} - \frac{1}{n^*}\right] \cdot \sum_{i=1}^{n^*} \xi_i^2 + \frac{1}{n} \cdot \sum_{i=n^*+1}^{n} \xi_i^2 + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \\
&= \frac{n^*-n}{n} \cdot \left(\frac{1}{n^*}\sum_{i=1}^{n^*} \xi_i^2\right) + \frac{n-n^*}{n} \cdot \left(\frac{1}{n-n^*}\sum_{i=n^*+1}^{n} \xi_i^2\right) + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \\
&= \frac{n^*-n}{n} \cdot \left(\sigma_0^2 + \mathrm{O}_{\mathbb{P}}\left(n^{*-\frac{1}{2}}\right)\right) + \frac{n-n^*}{n} \cdot \left(\sigma_0^2 + \mathrm{O}_{\mathbb{P}}\left((n-n^*)^{-\frac{1}{2}}\right)\right) + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),
\end{aligned}$$

using the Central Limit Theorem, and in the end we get

$$\begin{aligned}
\widehat{\sigma}_n^2 - \widehat{\sigma}_n^{2*} &= \frac{n^*-n}{n} \cdot \mathrm{O}_{\mathbb{P}}\left(n^{*-\frac{1}{2}}\right) + \frac{n-n^*}{n} \cdot \mathrm{O}_{\mathbb{P}}\left((n-n^*)^{-\frac{1}{2}}\right) + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \\
&= \mathrm{o}(1) \cdot \mathrm{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) + n^{-\frac{1}{2}} \cdot \mathrm{O}_{\mathbb{P}}\left(\left(\frac{n-n^*}{n}\right)^{\frac{1}{2}}\right) + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \\
&= \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) + n^{-\frac{1}{2}} \cdot \mathrm{O}_{\mathbb{P}}\left(\mathrm{o}(1)\right) + \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) = \mathrm{o}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

■

2.7 Technical results

Theorem 2.11 (Polya's Theorem). *Let $(g_n)_{n\in\mathbb{N}}$ be a sequence of non decreasing (or non increasing) functions defined over $I = [a, b] \subset \mathbb{R}$. If g_n converges pointwise to g (i.e. $g_n(x) \longrightarrow g(x)$ as $n \longrightarrow +\infty$, for any $x \in I$) and g is continuous then*

$$\sup_{x\in I} |g_n(x) - g(x)| \xrightarrow[n\longrightarrow+\infty]{} 0.$$

Proof of Lemma 2.11. Assume the functions g_n are non decreasing over I (if not, consider their opposites $-g_n$). g is continuous over I and thus bounded since I is compact. g is also non decreasing over I as the limit of a sequence of non decreasing functions. Let $\epsilon > 0$ and $k > \frac{g(b)-g(a)}{\epsilon}$ such that

$$\exists a = a_0 < \ldots < a_k = b \in I^{k+1},\ \forall i = 0, \ldots, k-1,\ g(a_{i+1}) - g(a_i) < \epsilon.$$

Now let $x \in I$ and let $i \in \mathbb{N}$ such that $a_i \leqslant x \leqslant a_{i+1}$. Since g_n and g are non decreasing, we find that

$$\begin{aligned}
g_n(x) - g(x) &\leqslant g_n(a_{i+1}) - g(a_i) \leqslant g_n(a_{i+1}) - g(a_{i+1}) + \epsilon, \\
g_n(x) - g(x) &\geqslant g_n(a_i) - g(a_{i+1}) \geqslant g_n(a_i) - g(a_i) - \epsilon.
\end{aligned}$$

The pointwise convergence of g_n to g and the finiteness of k together ensure that

$$\exists N_0 \in \mathbb{N},\ \forall n \geqslant N_0,\ \forall i = 0, \ldots, k,\ |g_n(a_i) - g(a_i)| < \epsilon,$$

which implies with both of the inequations mentioned above that

$$\exists N_0 \in \mathbb{N},\ \forall n \geqslant N_0,\ \forall x \in I,\ |g_n(x) - g(x)| < \epsilon.$$

■

Lemma 2.12. *Let $k \in \mathbb{N}^*$, there exists a constant $C \in \mathbb{R}_+^*$ such that for any $(u, u') \in [\underline{u},\, \overline{u}]^2$*

$$\sup_{t \in [\underline{u}, \overline{u}]} |(t-u')^k \mathbb{1}_{[t,+\infty[}(u') - (t-u)^k \mathbb{1}_{[t,+\infty[}(u)| = C|u-u'|. \tag{2.71}$$

Proof of Lemma 2.12. For any $(u, u') \in [\underline{u},\, \overline{u}]^2$ we have

$$\begin{aligned}\sup_{t \in [\underline{u}, \overline{u}]} |(t-u')^k \mathbb{1}_{[t,+\infty[}(u') - (t-u)^k \mathbb{1}_{[t,+\infty[}(u)| &\leqslant \sup_{t \in [\underline{u}, \overline{u}]} \{|(t-u')^k - (t-u)^k| \mathbb{1}_{[t,+\infty[}(u')\} \\ &\quad + \sup_{t \in [\underline{u}, \overline{u}]} \{|t-u|^k |\mathbb{1}_{[t,+\infty[}(u') - \mathbb{1}_{[t,+\infty[}(u)|\}. \end{aligned} \tag{2.72}$$

The mean value theorem guarantees that there exists v between u and u' such that

$$(t-u')^k - (t-u)^k = -k(t-v)^{k-1}(u'-u).$$

We thus have

$$\begin{aligned}\sup_{t \in [\underline{u}, \overline{u}]} \{|(t-u')^k - (t-u)^k| \mathbb{1}_{[t,+\infty[}(u')\} &\leqslant \sup_{t \in [\underline{u}, \overline{u}]} |(t-u')^k - (t-u)^k| \\ &\leqslant k|\overline{u} - \underline{u}|^{k-1}|u-u'|. \end{aligned} \tag{2.73}$$

Because $|t-u| \leqslant |u'-u|$ whenever $|\mathbb{1}_{[t,+\infty[}(u') - \mathbb{1}_{[t,+\infty[}(u)| \neq 0$, we also find that

$$\sup_{t \in [\underline{u}, \overline{u}]} \{|t-u|^k |\mathbb{1}_{[t,+\infty[}(u') - \mathbb{1}_{[t,+\infty[}(u)|\} \leqslant |u-u'|^k \leqslant |u-u'||\overline{u}-\underline{u}|^{k-1}. \tag{2.74}$$

And now (2.71) is a simple consequence of (2.72), (2.73) and (2.74). ■

Lemma 2.13. *For any $\eta' \in \mathbb{R} \times [\underline{u},\, \overline{u}]$, there exists $C \in \mathbb{R}_+^*$ such that for any $\eta \in \mathbb{R} \times [\underline{u},\, \overline{u}]$*

$$\sup_{t \in [\underline{u}, \overline{u}]} |\mu(\eta, t) - \mu(\eta', t)| \leqslant C\|\eta - \eta'\|. \tag{2.75}$$

Proof of Lemma 2.13. We have indeed

$$\begin{aligned}\sup_{t \in [\underline{u}, \overline{u}]} |\mu(\eta, t) - \mu(\eta', t)| &= \sup_{t \in [\underline{u}, \overline{u}]} |\gamma \cdot (t-u)\mathbb{1}_{[t,+\infty[}(u) - \gamma'(t-u')\mathbb{1}_{[t,+\infty[}(u')| \\ &\leqslant \sup_{t \in [\underline{u}, \overline{u}]} |[\gamma - \gamma'](t-u)\mathbb{1}_{[t,+\infty[}(u)| \\ &\qquad + \sup_{t \in [\underline{u}, \overline{u}]} |\gamma'[(t-u)\mathbb{1}_{[t,+\infty[}(u) - (t-u')\mathbb{1}_{[t,+\infty[}(u')]| \\ &\leqslant |\gamma - \gamma'| \cdot \sup_{t \in [\underline{u}, \overline{u}]} |t-u| \\ &\qquad + |\gamma'| \cdot \sup_{t \in [\underline{u}, \overline{u}]} |(t-u)\mathbb{1}_{[t,+\infty[}(u) - (t-u')\mathbb{1}_{[t,+\infty[}(u')| \\ &\leqslant |\gamma - \gamma'| \cdot |\overline{u} - \underline{u}| + |\gamma'| \cdot \sup_{t \in [\underline{u}, \overline{u}]} |(t-u)\mathbb{1}_{[t,+\infty[}(u) - (t-u')\mathbb{1}_{[t,+\infty[}(u')|. \end{aligned}$$

And now (2.75) is a simple consequence of Lemma 2.12. ■

Lemma 2.14. *Let $A \subset \mathbb{R} \times [\underline{u}, \overline{u}]$ be a bounded set. Then,*

$$\forall \epsilon > 0,\ \exists m(\epsilon) \in \mathbb{N},\ \exists \eta_1, \ldots, \eta_{m(\epsilon)} \in A,$$
$$\forall \eta, \eta' \in A,\ \exists j, j' \in \{1, \ldots, m(\epsilon)\},\ \sup_{t \in [\underline{u}, \overline{u}]} \Big| \big[\mu(\eta, t) - \mu(\eta', t)\big] - \big[\mu(\eta_j, t) - \mu(\eta_{j'}, t)\big] \Big| < \epsilon,$$

Proof of Lemma 2.14. It suffices to prove the following claim

$$\forall \epsilon > 0,\ \exists m(\epsilon) \in \mathbb{N},\ \exists \eta_1, \ldots, \eta_{m(\epsilon)} \in A,$$
$$\forall \eta \in A,\ \exists j \in \{1, \ldots, m(\epsilon)\},\ \sup_{t \in [\underline{u}, \overline{u}]} |\mu(\eta, t) - \mu(\eta_j, t)| < \epsilon.$$

and then use the triangle inequality. To see that the claim holds, it suffices, thanks to Lemma 2.13, to exhibit a finite and tight enough grid of A such that any point in A lies close enough to a point of the grid. The existence of such a grid is obviously guaranteed since $A \subset \mathbb{R}^2$ is bounded. ■

Lemma 2.15. *Recall the definition of b_n given in* (2.24). *Let*

$$b(\theta) = \left(\frac{\sigma_0^2}{\sigma^2} - 1 - \log \frac{\sigma_0^2}{\sigma^2} \right) + \frac{1}{\sigma^2} \int_{\underline{u}}^{\overline{u}} \left[\mu(\eta_0, t) - \mu(\eta, t)\right]^2 f(t)\, \mathrm{d}t. \tag{2.76}$$

Then, under Assumptions A1–A4,

$$b_n(\theta) \geqslant 0. \tag{2.77}$$
$$b(\theta) \geqslant 0, \text{ with equality if and only if } \theta = \theta_0. \tag{2.78}$$
$$b_n(\theta') \to b_n(\theta), \text{ uniformly in } n, \text{ as } \theta' \to \theta. \tag{2.79}$$
$$b_n(\theta) \to b(\theta), \text{ as } n \to +\infty. \tag{2.80}$$

Proof of Lemma 2.15. We will prove each claim separately.

Proof of (2.77). That $b_n(\theta) \geqslant 0$ is trivial since the first term in (2.24) is non negative (having $x - 1 - \log x \geqslant 0$ with equality only if $x = 1$), and the second term in (2.24) is obviously non negative too.

Proof of (2.78). That $b(\theta) \geqslant 0$ is again easy enough to prove, both terms in (2.76) being trivially non negative. If $\theta \neq \theta_0$ then either $\sigma^2 \neq \sigma_0^2$ which implies the first term is positive, or $\mu(\eta_0, \cdot) \neq \mu(\eta, \cdot)$ which implies the second term is positive (since f is assumed positive on $[\underline{u}, \overline{u}]$). Hence if $\theta \neq \theta_0$ then $b(\theta) > 0$. That $\theta = \theta_0$ implies $b(\theta) = 0$ is of course straightforward.

Proof of (2.79). We first observe that

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^n (\mu(\eta_0, t_i) - \mu(\eta', t_i))^2 - \frac{1}{n} \sum_{i=1}^n (\mu(\eta_0, t_i) - \mu(\eta, t_i))^2 \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n \left[2\mu(\eta_0, t_i) - \mu(\eta', t_i) - \mu(\eta, t_i)\right] \cdot \left[\mu(\eta, t_i) - \mu(\eta', t_i)\right] \right| \\
&\leqslant \frac{1}{n} \sum_{i=1}^n \left|2\mu(\eta_0, t_i) - \mu(\eta', t_i) - \mu(\eta, t_i)\right| \cdot \left|\mu(\eta, t_i) - \mu(\eta', t_i)\right| \\
&\leqslant \left(\sup_{t \in [\underline{u}, \overline{u}]} |\mu(\eta_0, t) - \mu(\eta, t)| + \sup_{t \in [\underline{u}, \overline{u}]} \left|\mu(\eta', t) - \mu(\eta, t)\right| \right) \cdot \sup_{t \in [\underline{u}, \overline{u}]} \left|\mu(\eta', t) - \mu(\eta, t)\right|.
\end{aligned} \tag{2.81}$$

As $\theta' \to \theta$, the convergence of the first term of b_n to the first term of b is obviously uniform in n since this part of b_n does not involve n at all. As $\theta' \to \theta$, via Lemma 2.13, we also obtain

$$\sup_{t\in[\underline{u},\overline{u}]} \left|\mu(\eta',t) - \mu(\eta,t)\right| \to 0,$$

which ensures that the second part of (2.24) converges uniformly in n thanks to (2.81).

Proof of (2.80). Thanks to Assumption A1, it is easy to see that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_0,t_i) - \mu(\eta,t_i)]^2 &= \int_{\underline{u}}^{\overline{u}} [\mu(\eta_0,t) - \mu(\eta,t)]^2 \,\mathrm{d}F_n(t) \\ &\to \int_{\underline{u}}^{\overline{u}} [\mu(\eta_0,t) - \mu(\eta,t)]^2 \,\mathrm{d}F(t) = \int_{\underline{u}}^{\overline{u}} [\mu(\eta_0,t) - \mu(\eta,t)]^2 f(t)\,\mathrm{d}t.\end{aligned}$$

■

Lemma 2.16. *Let $A \subset \mathbb{R} \times [\underline{u}, \overline{u}]$ be a bounded set, and let $\eta_0 \in A$, then under Assumptions A1–A4,*

$$\sup_{\eta\in A}\left|\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_0,t_i) - \mu(\eta,t_i)]\xi_i\right| \xrightarrow{a.s} 0.$$

Proof of Lemma 2.16. Let $\epsilon > 0$, $\eta \in A$, and apply Lemma 2.14 to get the corresponding $m(\epsilon) \in \mathbb{N}$, $\{\eta_1,\ldots,\eta_{m(\epsilon)}\} \subset A, j, j' \in \{1,\ldots,m(\epsilon)\}$. We can write with the triangle inequality

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_0,t_i) - \mu(\eta,t_i)]\xi_i &= \frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_j,t_i) - \mu(\eta_{j'},t_i)]\xi_i \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\Big\{[\mu(\eta_0,t_i) - \mu(\eta,t_i)] - [\mu(\eta_j,t_i) - \mu(\eta_{j'},t_i)]\Big\}\xi_i \\ &\leqslant \sup_{(j,j')\in\{1,\ldots,m(\epsilon)\}}\left\{\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_j,t_i) - \mu(\eta_{j'},t_i)]\xi_i\right\} + \epsilon\cdot\frac{1}{n}\sum_{i=1}^{n}|\xi_i|\,.\end{aligned}$$

Hence

$$\begin{aligned}\sup_{\eta\in A}\left\{\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_0,t_i) - \mu(\eta,t_i)]\xi_i\right\} &\leqslant \sup_{(j,j')\in\{1,\ldots,m(\epsilon)\}}\left\{\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_j,t_i) - \mu(\eta_{j'},t_i)]\xi_i\right\} \\ &\quad + \epsilon\cdot\frac{1}{n}\sum_{i=1}^{n}|\xi_i|\,. \end{aligned}\tag{2.82}$$

Let us now recall Kolmogorov's criterion, a proof of which is available in Section 17 of Loève (1991) on pages 250–251. This criterion guarantees that for any sequence $(Y_i)_{i\in\mathbb{N}}$ of independent random variables and any numerical sequence $(b_i)_{i\in\mathbb{N}}$ such that

$$\sum_{i=1}^{+\infty}\frac{\operatorname{Var} Y_i}{b_i^2} < +\infty,\ b_n \to +\infty,$$

we have

$$\frac{\sum_{i=1}^{n}(Y_i - \mathbb{E}Y_i)}{b_n} \xrightarrow{a.s} 0.$$

For each couple $(j, j') \in \{1, \dots, m(\epsilon)\}$, Kolmogorov's criterion ensures that

$$\frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_j, t_i) - \mu(\eta_{j'}, t_i)]\xi_i \xrightarrow{a.s} 0,$$

for the coefficients $[\mu(\eta_j, t_i) - \mu(\eta_{j'}, t_i)]$ are obviously bounded, and it suffices to pick $Y_i = [\mu(\eta_j, t_i) - \mu(\eta_{j'}, t_i)]\xi_i$ and $b_i = i$. Having only a finite number of couples $(j, j') \in \{1, \dots, m(\epsilon)\}^2$ to consider allows us to write

$$\sup_{(j,j')\in\{1,\dots,m(\epsilon)\}} \frac{1}{n}\sum_{i=1}^{n}[\mu(\eta_j, t_i) - \mu(\eta_{j'}, t_i)]\xi_i \xrightarrow{a.s} 0. \tag{2.83}$$

By (2.83), the first term on the right hand side of (2.82) converges almost surely to zero. The Strong Law of Large Numbers ensures that the second term on the right hand side of (2.82) converges almost surely to $\epsilon \cdot (2\pi^{-1}\sigma^2)^{\frac{1}{2}}$, and the result follows, since all the work done above for $(\xi_n)_{n\in\mathbb{N}}$ can be done again for $(-\xi_n)_{n\in\mathbb{N}}$. ■

Lemma 2.17. *Let $(Z_i)_{i\in\mathbb{N}}$ be a sequence of independent identically distributed random variables such that for all $i \in \mathbb{N}$, either $Z_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$, or $Z_i \sim \chi^2(k)$ with $k > 0$. Then a.s., as $n \to +\infty$*

$$Z_n = \mathrm{O}(\log n).$$

Proof of Lemma 2.17. Denote $Y_n = Z_n$ when the random variables are Gaussian, and $Y_n = Z_n/5$ when the random variables considered are chi-squared (so that $\mathbb{E}e^{2Y_1}$ and $\mathbb{E}e^{-2Y_1}$ are both finite). We will show that a.s. $Y_n = \mathrm{O}(\log n)$.

For any $\epsilon > 0$, from Markov's inequality we get:

$$\mathbb{P}\left(n^{-1}|e^{Y_n}| > \epsilon\right) = \mathbb{P}\left(n^{-2}e^{2Y_n} > \epsilon^2\right) \leqslant \epsilon^{-2}n^{-2}\mathbb{E}e^{2Y_1}.$$

From there it is easy to see that for any $\epsilon > 0$ we have

$$\sum_{n=1}^{+\infty}\mathbb{P}\left(n^{-1}|e^{Y_n}| > \epsilon\right) = \epsilon^{-2}\frac{\pi^2}{6}\mathbb{E}e^{2Y_1} < \infty,$$

which directly implies via Borel-Cantelli's Lemma (see for example Billingsley, 1995, Section 4, page 59) that a.s.

$$e^{Y_n} = \mathrm{o}(n).$$

In particular, a.s. for any n large enough,

$$Y_n \leqslant \log n.$$

What was done with $(Y_n)_{n\in\mathbb{N}}$ can be done again with $(-Y_n)_{n\in\mathbb{N}}$ so that in the end we have a.s for any n large enough,

$$-\log n \leqslant Y_n \leqslant \log n.$$

■

Lemma 2.18. *Under Assumptions A1–A4, for any $\eta_0 \in \mathbb{R} \times [\underline{u}, \overline{u}]$, there exists $C \in \mathbb{R}_+^*$ such that for any n large enough, and for any η*

$$n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 \geqslant C\|\eta - \eta_0\|^2.$$

Proof of Lemma 2.18. We have already almost proved this result in (2.18) (see Theorem 2.7). There is however a small difficulty since the majoration was obtained for $\tau = (\beta, \gamma)$ and not $\eta = (\gamma, u)$.

Let V_1 and V_2 two non empty open intervals of $]\underline{u},\, u_0[$ such that their closures $\overline{V_1}$ and $\overline{V_2}$ are do not overlap. We have

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 \geqslant n^{-1} \Bigg(& \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2\, \mathbb{1}_{V_1}(t_i) + \\ & \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2\, \mathbb{1}_{V_2}(t_i) \Bigg). \end{aligned}$$

Using the same arguments we used to prove (2.18), we find that there exists $C \in \mathbb{R}_+^*$ such that (remembering the definition of the intercept β of the model)

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 &\geqslant \min \left(n^{-1} \sum_{i=1}^{n} \mathbb{1}_{V_1}(t_i), n^{-1} \sum_{i=1}^{n} \mathbb{1}_{V_2}(t_i) \right) \cdot C|\gamma - \gamma_0|^2, \\ n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 &\geqslant \min \left(n^{-1} \sum_{i=1}^{n} \mathbb{1}_{V_1}(t_i), n^{-1} \sum_{i=1}^{n} \mathbb{1}_{V_2}(t_i) \right) \cdot C|\beta - \beta_0|^2, \end{aligned}$$

and since for $j = 1, 2$ we have

$$n^{-1} \sum_{i=1}^{n} \mathbb{1}_{V_j}(t_i) \to \int_{V_j} f(t)\, \mathrm{d}t > 0,$$

there exists $C \in \mathbb{R}_+^*$ such that for any n large enough

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 &\geqslant C|\gamma - \gamma_0|^2, \\ n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 &\geqslant C|\beta - \beta_0|^2. \end{aligned}$$

Notice now that

$$\begin{aligned} |u - u_0| &= |\gamma_0^{-1}\beta_0 - \gamma^{-1}\beta| \\ &= |\gamma_0^{-1}||\beta_0 - \gamma_0\gamma^{-1}\beta| \\ &\leqslant |\gamma_0^{-1}| \left\{ |\beta_0 - \beta| + |\beta - \gamma_0\gamma^{-1}\beta| \right\} \\ &\leqslant |\gamma_0^{-1}| \left\{ |\beta_0 - \beta| + |\gamma^{-1}\beta|\, |\gamma - \gamma_0| \right\} \\ &\leqslant |\gamma_0^{-1}|(|\beta_0 - \beta| + |u|\, |\gamma - \gamma_0|) \\ &\leqslant |\gamma_0^{-1}|(1 + \max(|\underline{u}|, |\overline{u}|)) \cdot \max(|\beta_0 - \beta|, |\gamma - \gamma_0|). \end{aligned}$$

From here, since $u \in [\underline{u}, \overline{u}]$ is bounded, it is straightforward that there exists $C \in \mathbb{R}_+^*$ such that for any n large enough

$$n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 \geqslant C|\gamma - \gamma_0|^2,$$
$$n^{-1} \sum_{i=1}^{n} [\mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 \geqslant C|u - u_0|^2,$$

which ends the proof. ∎

Lemma 2.19. *Recall the definition of $A^*_{1:n}$ given in* (2.68). *Under Assumptions A1–A4 and conditions* (2.10)*, as $n \to +\infty$*

$$n^{-\frac{1}{2}} A^*_{1:n}(\theta_0) \xrightarrow{d} \mathcal{N}\left(0, I(\theta_0)\right). \tag{2.84}$$

Proof of Lemma 2.19. We will show that any linear combination of the coordinates of $A_{1:n}(\theta_0)$ is asymptotically normal using Lyapounov's Theorem. Let $\alpha \in \mathbb{R}^3$, $\|\alpha\| \neq 0$, so that differential calculus allows us to write

$$\begin{aligned}
\langle \alpha, A^*_{1:n}(\theta_0) \rangle &= \alpha_1 \cdot \left.\frac{\partial l^*_{1:n}(X_{1:n}|\theta)}{\partial \gamma}\right|_{\theta_0} + \alpha_2 \cdot \left.\frac{\partial l^*_{1:n}(X_{1:n}|\theta)}{\partial u}\right|_{\theta_0} + \alpha_3 \cdot \left.\frac{\partial l^*_{1:n}(X_{1:n}|\theta)}{\partial \sigma^2}\right|_{\theta_0} \\
&= \alpha_1 \cdot \frac{1}{\sigma_0^2} \sum_{i=1}^{n^*} \left[(t_i - u_0) \mathbb{1}_{[t_i, +\infty[}(u_0) \cdot \xi_i \right] - \alpha_2 \cdot \frac{\gamma_0}{\sigma_0^2} \sum_{i=1}^{n^*} \left[\mathbb{1}_{[t_i, +\infty[}(u_0) \cdot \xi_i \right] \\
&\quad + \alpha_3 \cdot \frac{1}{2\sigma_0^2} \sum_{i=1}^{n^*} \left[\frac{1}{\sigma_0^2} \cdot \xi_i^2 - 1 \right] \\
&= \sigma_0^{-2} \sum_{i=1}^{n^*} Z_i,
\end{aligned}$$

where we denote, for $i = 1, \ldots, n^*$

$$Z_i = \left[\left\{ (t_i - u_0) \mathbb{1}_{[t_i, +\infty[}(u_0) \cdot \alpha_1 - \gamma_0 \mathbb{1}_{[t_i, +\infty[}(u_0) \cdot \alpha_2 \right\} \cdot \xi_i + \frac{1}{2} \alpha_3 \cdot \left\{ \sigma_0^{-2} \cdot \xi_i^2 - 1 \right\} \right]. \tag{2.85}$$

Since for $i = 1, \ldots, n^*$ $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i^2] = \sigma^2$, we deduce that $\mathbb{E}[Z_i] = 0$, and hence that $\mathbb{E}\left[\langle \alpha, A^*_{1:n}(\theta_0) \rangle\right] = 0$.

Let us now find the expression of $\operatorname{Var} \langle \alpha, A^*_{1:n}(\theta_0) \rangle$. Because ξ_i and ξ_j are independent when $i \neq j$, so are Z_i and Z_j and we hence write

$$\begin{aligned}
\operatorname{Var} \langle \alpha, A^*_{1:n}(\theta_0) \rangle &= \sigma_0^{-4} \sum_{i=1}^{n^*} \operatorname{Var} Z_i \\
&= \sigma_0^{-4} \sum_{i=1}^{n^*} \left\{ \left[(t_i - u_0) \mathbb{1}_{[t_i, +\infty[}(u_0) \cdot \alpha_1 - \gamma_0 \mathbb{1}_{[t_i, +\infty[}(u_0) \cdot \alpha_2 \right]^2 \cdot \operatorname{Var} \xi_i \right. \\
&\qquad \left. + \frac{1}{4} \alpha_3^2 \cdot \operatorname{Var} \left[\sigma_0^{-2} \xi_i^2 - 1 \right] \right\},
\end{aligned}$$

because Cov $\left[\xi_i, \left\{\sigma_0^{-2}\xi_i^2 - 1\right\}\right] = 0$, and we finally get

$$\operatorname{Var}\left\langle \alpha, A^*_{1:n}(\theta_0)\right\rangle = \sigma_0^{-4}\sum_{i=1}^{n^*}\left\{\left[(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_1 - \gamma_0\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_2\right]^2\cdot\sigma_0^2 + \frac{1}{4}\alpha_3^2\cdot 2\right\}.$$

We can hence write

$$\begin{aligned}
n^{*-1}\operatorname{Var}\left\langle \alpha, A^*_{1:n}(\theta_0)\right\rangle &= \sigma_0^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_1 - \gamma_0\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_2\right]^2 + \frac{1}{2}\sigma_0^{-4}\alpha_3^2\\
&= \alpha_1^2\cdot\sigma_0^{-2}\left\{\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i - u_0)^2\mathbb{1}_{[t_i,+\infty[}(u_0)\right\}\\
&\quad - 2\alpha_1\alpha_2\cdot\sigma_0^{-2}\gamma_0\left\{\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i - u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\right\}\\
&\quad + \alpha_2^2\cdot\sigma_0^{-2}\gamma_0^2\left\{\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u_0)\right\} + \alpha_3^2\cdot\frac{1}{2}\sigma_0^{-4}\\
&= \left\langle \alpha, I_{1:n}(\theta_0)\alpha\right\rangle,
\end{aligned}$$

where we denote

$$I^*_{1:n}(\theta) = \begin{bmatrix} \sigma^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i - u)^2\mathbb{1}_{[t_i,+\infty[}(u) & -\sigma^{-2}\gamma\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i - u)\mathbb{1}_{[t_i,+\infty[}(u) & 0\\ & \sigma^{-2}\gamma^2\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u) & 0\\ & & \frac{1}{2}\sigma^{-4}\end{bmatrix}. \qquad (2.86)$$

Remark that, by virtue of Assumption A1, it is easy to check that for any $\theta \in \Theta$

$$I^*_{1:n}(\theta) \to I(\theta), \qquad (2.87)$$

and observe that just like $I(\theta)$, $I^*_{1:n}(\theta)$ is positive definite, since all its principal minor determinants are positive.

Let us now check that the random variables Z_i meet Lyapounov's Theorem (see Billingsley, 1995, page 362) requirements before wrapping up this proof. The random variables Z_i are independent and trivially L^2. We denote $V_n^{*2} = \sum_{i=1}^{n^*}\operatorname{Var} Z_i$ and claim that Lyapounov's condition holds, that is

$$\exists\delta > 0,\ \sum_{i=1}^{n^*}\mathbb{E}\left|\frac{Z_i - \mathbb{E}Z_i}{V_n^*}\right|^{2+\delta} = \mathrm{o}(1).$$

Indeed we have ($\delta = 1$)

$$\begin{aligned}
\sum_{i=1}^{n^*}\mathbb{E}\left|\frac{Z_i - \mathbb{E}Z_i}{V_n^*}\right|^3 &= \sum_{i=1}^{n^*}\mathbb{E}\left|\frac{Z_i}{V_n^*}\right|^3\\
&= \frac{n^*}{\operatorname{Var}^{\frac{3}{2}}\left\langle \alpha, A^*_{1:n}(\theta_0)\right\rangle}\cdot\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{E}\left|Z_i\right|^3\\
&= \frac{1}{n^{*\frac{1}{2}}\left\langle \alpha, I^*_{1:n}(\theta_0)\alpha\right\rangle^{\frac{3}{2}}}\cdot\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{E}\left|Z_i\right|^3.
\end{aligned}$$

The first term of this last product is $\mathrm{O}\left(n^{*-\frac{1}{2}}\right)$ thanks to (2.87), and recalling the definition of Z_i from (2.85), there is no difficulty in showing that the last term of the product, namely $\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{E}\left|Z_i\right|^3$ converges to a finite limit. Indeed we find, using trivial dominations and Assumption A1 once again,

$$\begin{aligned}
|Z_i|^3 &= \left|\left\{(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_1-\gamma_0\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_2\right\}\cdot\xi_i+\frac{1}{2}\alpha_3\cdot\left\{\sigma_0^{-2}\cdot\xi_i^2-1\right\}\right|^3\\
\mathbb{E}|Z_i|^3 &\leqslant \left(\left|(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_1-\gamma_0\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_2\right|+\left|\frac{1}{2}\alpha_3\right|\right)^3\\
&\qquad\times\mathbb{E}\left(|\xi_i|+\left|\sigma_0^{-2}\cdot\xi_i^2-1\right|\right)^3\\
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|Z_i|^3 &\leqslant \frac{1}{n}\sum_{i=1}^{n}\left(\left|(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_1-\gamma_0\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_2\right|+\left|\frac{1}{2}\alpha_3\right|\right)^3\\
&\qquad\times\mathbb{E}\left(|\xi_i|+\left|\sigma_0^{-2}\cdot\xi_i^2-1\right|\right)^3\\
&\leqslant \mathrm{O}(1)\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\left|(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_1-\gamma_0\mathbb{1}_{[t_i,+\infty[}(u_0)\cdot\alpha_2\right|+\left|\frac{1}{2}\alpha_3\right|\right)^3\\
&\leqslant \mathrm{O}(1).
\end{aligned}$$

Lyapounov's Theorem thus applies here and leads to

$$\sum_{i=1}^{n^*}\frac{Z_i-\mathbb{E}Z_i}{V_n^*}\xrightarrow{d}\mathcal{N}(0,1),$$

i.e. multiplying numerator and denominator by σ_0^{-2} we get

$$\frac{\left\langle\alpha,A_{1:n}^*(\theta_0)\right\rangle}{\mathrm{Var}^{\frac{1}{2}}\left\langle\alpha,A_{1:n}^*(\theta_0)\right\rangle}\xrightarrow{d}\mathcal{N}(0,1),$$

that is

$$\frac{\left\langle\alpha,A_{1:n}^*(\theta_0)\right\rangle}{n^{*\frac{1}{2}}\left\langle\alpha,I_{1:n}^*(\theta_0)\alpha\right\rangle^{\frac{1}{2}}}\xrightarrow{d}\mathcal{N}(0,1),$$

and because of (2.87) we can also write,

$$\frac{\left\langle\alpha,A_{1:n}^*(\theta_0)\right\rangle}{n^{*\frac{1}{2}}\left\langle\alpha,I(\theta_0)\alpha\right\rangle^{\frac{1}{2}}}\xrightarrow{d}\mathcal{N}(0,1),$$

which, remembering that a.s. $n^*\sim n$, is equivalent to (2.84). ∎

Lemma 2.20. *Recall the definition of $B_{1:n}^*$ given in* (2.36). *Under Assumptions A1–A4 and conditions* (2.10), *as $n\to+\infty$,*

$$\frac{1}{n}B_{1:n}^*(\theta_0)\xrightarrow{a.s}I(\theta_0),\ \text{as } n\to+\infty. \tag{2.88}$$

$$\frac{1}{n}B_{1:n}^*(\theta)\xrightarrow{a.s}I(\theta_0),\ \text{as } \theta\to\theta_0 \text{ and } n\to+\infty. \tag{2.89}$$

where the asymptotic Fisher Information Matrix $I(\cdot)$ is defined in (2.7).

Proof of Lemma 2.20. We will prove each claim separately.

Proof of (2.88). Differential calculus provides the following expressions for the coefficients of $\frac{1}{n^*}B^*_{1:n}(\theta)$.

$$\begin{aligned}
\left(\frac{1}{n^*}B^*_{1:n}(\theta)\right)_{11} &= \sigma^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i-u)^2\mathbb{1}_{[t_i,+\infty[}(u),\\
\left(\frac{1}{n^*}B^*_{1:n}(\theta)\right)_{12} &= \sigma^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}\Big[\xi_i+\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-2\gamma\cdot(t_i-u)\Big]\,\mathbb{1}_{[t_i,+\infty[}(u),\\
\left(\frac{1}{n^*}B^*_{1:n}(\theta)\right)_{13} &= \sigma^{-4}\frac{1}{n^*}\sum_{i=1}^{n^*}\Big[\xi_i+\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-\gamma\cdot(t_i-u)\Big]\,(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u),\\
\left(\frac{1}{n^*}B^*_{1:n}(\theta)\right)_{22} &= \sigma^{-2}\gamma^2\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u),\\
\left(\frac{1}{n^*}B^*_{1:n}(\theta)\right)_{23} &= -\sigma^{-4}\gamma\frac{1}{n^*}\sum_{i=1}^{n^*}\Big[\xi_i+\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-\gamma\cdot(t_i-u)\Big]\,\mathbb{1}_{[t_i,+\infty[}(u),\\
\left(\frac{1}{n^*}B^*_{1:n}(\theta)\right)_{33} &= -\frac{1}{2}\sigma^{-4}+\sigma^{-6}\frac{1}{n^*}\sum_{i=1}^{n^*}\Big[\xi_i+\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-\gamma\cdot(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u)\Big]^2.
\end{aligned}$$

The convergence we claim is then a direct consequence of Assumption A1 and the fact that $n^*\sim n$ and, depending on the coefficients, either the Strong Law of Large Numbers or Kolmogorov's criterion. Notice that

$$\frac{1}{n^*}B^*_{1:n}(\theta_0)-I^*_{1:n}(\theta_0)\xrightarrow{a.s}0,$$

where $I^*_{1:n}$ is defined in (2.86).

Proof of (2.89). We will show that in fact, as $n\to+\infty$ and $\theta\to\theta_0$,

$$C^*_{1:n}(\theta)=\frac{1}{n^*}B^*_{1:n}(\theta_0)-\frac{1}{n^*}B^*_{1:n}(\theta)\xrightarrow{a.s}0,$$

which will end the proof since $n^*\sim n$. We will consider each coefficient of $C^*_{1:n}(\theta)$ in turn, making use of Assumption A1 once again and apply repeatedly the Strong Law of Large Numbers and Kolmogorov's criterion as well as Lemma 2.12, whenever needed.

$$\begin{aligned}
C^*_{1:n}(\theta)_{11} &= \sigma_0^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i-u_0)^2\mathbb{1}_{[t_i,+\infty[}(u_0)-\sigma^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i-u)^2\mathbb{1}_{[t_i,+\infty[}(u)\\
&= \left(\sigma_0^{-2}-\sigma^{-2}\right)\cdot\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i-u_0)^2\mathbb{1}_{[t_i,+\infty[}(u_0)\\
&\quad+\sigma^{-2}\cdot\left(\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i-u_0)^2\mathbb{1}_{[t_i,+\infty[}(u_0)-\frac{1}{n^*}\sum_{i=1}^{n^*}(t_i-u)^2\mathbb{1}_{[t_i,+\infty[}(u)\right)\\
&= \mathrm{o}(1)\cdot\mathrm{O}(1)+\mathrm{O}(1)\cdot\mathrm{O}\left(u-u_0\right)\to 0.
\end{aligned}$$

then last equality holding true because of Lemma 2.12.

$$\begin{aligned}
C^*_{1:n}(\theta)_{22} &= \sigma_0^{-2}\gamma_0^2\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u_0) - \sigma^{-2}\gamma^2\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u) \\
&= \left(\sigma_0^{-2}\gamma_0^2 - \sigma^{-2}\gamma^2\right)\cdot\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u_0) \\
&\quad + \sigma^{-2}\gamma^2\cdot\left[\frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u_0) - \frac{1}{n^*}\sum_{i=1}^{n^*}\mathbb{1}_{[t_i,+\infty[}(u)\right] \\
&= \mathrm{o}(1)\cdot\mathrm{O}(1) + \mathrm{O}(1)\cdot\left[\{F_{n^*}(u_0) - F(u_0)\} + \{F(u_0) - F(u)\} + \{F(u) - F_{n^*}(u)\}\right] \\
&= \mathrm{o}(1) + \mathrm{O}(1)\cdot[\mathrm{o}(1) + \mathrm{o}(1) + \mathrm{o}(1)] \to 0,
\end{aligned}$$

the last equality holding true because of the uniform convergence of F_{n^*} to F over any compact subset such as $[\underline{u}, \overline{u}]$ (see Assumption A1, and its Remark 1).

$$\begin{aligned}
C^*_{1:n}(\theta)_{33} &= \frac{1}{2}\sigma^{-4} - \sigma^{-6}\frac{1}{n^*}\sum_{i=1}^{n^*}[\xi_i + \gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \gamma\cdot(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u)]^2 \\
&\quad - \left(\frac{1}{2}\sigma_0^{-4} - \sigma_0^{-6}\frac{1}{n^*}\sum_{i=1}^{n^*}\xi_i^2\right) \\
&= \frac{1}{2}(\sigma^{-4} - \sigma_0^{-4}) - (\sigma^{-6} - \sigma_0^{-6})\cdot\frac{1}{n^*}\sum_{i=1}^{n^*}\xi_i^2 \\
&\quad - \sigma^{-6}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \gamma\cdot(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u)\right]\xi_i \\
&\quad - \sigma^{-6}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \gamma\cdot(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u)\right]^2 \\
&= \mathrm{o}(1) + \mathrm{o}(1)\cdot\frac{1}{n^*}\sum_{i=1}^{n^*}\xi_i^2 + \mathrm{o}(1) + \mathrm{o}(1) \xrightarrow{a.s} 0,
\end{aligned}$$

where the two last $\mathrm{o}(1)$ are direct consequences of Lemmas 2.13 and 2.16. Those same Lemmas used together with Lemma 2.12, the Strong Law of Large Numbers as well as the well-known Cauchy-Schwarz inequality imply that a.s.

$$\begin{aligned}
C^*_{1:n}(\theta)_{23} &= \sigma^{-4}\gamma\frac{1}{n^*}\sum_{i=1}^{n^*}\left[\xi_i + \gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \gamma\cdot(t_i-u)\right]\mathbb{1}_{[t_i,+\infty[}(u) \\
&\quad - \sigma_0^{-4}\gamma_0\frac{1}{n^*}\sum_{i=1}^{n^*}\xi_i\mathbb{1}_{[t_i,+\infty[}(u_0) \\
&= \frac{1}{n^*}\sum_{i=1}^{n^*}\left[\sigma^{-4}\gamma\mathbb{1}_{[t_i,+\infty[}(u) - \sigma_0^{-4}\gamma_0\mathbb{1}_{[t_i,+\infty[}(u_0)\right]\xi_i \\
&\quad + \frac{1}{n^*}\sum_{i=1}^{n^*}\left[\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) - \gamma\cdot(t_i-u)\right]\sigma^{-4}\gamma\mathbb{1}_{[t_i,+\infty[}(u) \\
&= \mathrm{o}(1) + \mathrm{o}(1) \xrightarrow{a.s} 0,
\end{aligned}$$

and also that a.s.

$$\begin{aligned}
C^*_{1:n}(\theta)_{13} &= \sigma_0^{-4}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)\right]\xi_i \\
&\quad -\sigma^{-4}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[\xi_i+\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-\gamma\cdot(t_i-u)\right](t_i-u)\mathbb{1}_{[t_i,+\infty[}(u) \\
&= \frac{1}{n^*}\sum_{i=1}^{n^*}\left[\sigma_0^{-4}(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-\sigma^{-4}(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u)\right]\xi_i \\
&\quad -\sigma^{-4}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-\gamma\cdot(t_i-u)\right](t_i-u)\mathbb{1}_{[t_i,+\infty[}(u) \\
&= \mathrm{o}(1)+\mathrm{o}(1)\xrightarrow{a.s}0.
\end{aligned}$$

and finally that a.s.

$$\begin{aligned}
C^*_{1:n}(\theta)_{12} &= \sigma_0^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[\xi_i-\gamma_0\cdot(t_i-u_0)\right]\mathbb{1}_{[t_i,+\infty[}(u_0) \\
&\quad -\sigma^{-2}\frac{1}{n^*}\sum_{i=1}^{n^*}\left[\xi_i+\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-2\gamma\cdot(t_i-u)\right]\mathbb{1}_{[t_i,+\infty[}(u) \\
&= \frac{1}{n^*}\sum_{i=1}^{n^*}\xi_i\cdot\left[\sigma_0^{-2}\mathbb{1}_{[t_i,+\infty[}(u_0)-\sigma^{-2}\mathbb{1}_{[t_i,+\infty[}(u)\right] \\
&\quad +\frac{1}{n^*}\sum_{i=1}^{n^*}\Big[-\sigma_0^{-2}\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0)-\sigma^{-2}(\gamma_0\cdot(t_i-u_0)\mathbb{1}_{[t_i,+\infty[}(u_0) \\
&\qquad\qquad -2\gamma\cdot(t_i-u)\mathbb{1}_{[t_i,+\infty[}(u))\Big] \\
&= \mathrm{o}(1)+\mathrm{o}(1)\xrightarrow{a.s}0.
\end{aligned}$$

■

Proposition 2.21. *Let* $0<\delta$, *and let* $(\rho_n)_{n\in\mathbb{N}}$ *be a positive sequence such that, as* $n\to+\infty$

$$\rho_n=\mathrm{O}(1) \tag{2.90}$$

$$n^{-\frac{1}{2}}(\log n)\cdot\rho_n^{-1}\to 0 \tag{2.91}$$

and denote

$$B^c(\theta_0,\delta\rho_n)=\{\theta\in\Theta,\ \|\theta-\theta_0\|\geqslant\delta\rho_n\},$$

Then, under Assumptions A1–A4, a.s., there exists $\epsilon>0$ *such that, for any n large enough*

$$\sup_{\theta\in B^c(\theta_0,\delta\rho_n)}\frac{1}{n\rho_n^2}\left[l_{1:n}(X_{1:n}|\theta)-l_{1:n}(X_{1:n}|\widehat{\theta}_n)\right]\leqslant-\epsilon. \tag{2.92}$$

$$\sup_{\theta\in B^c(\theta_0,\delta\rho_n)}\frac{1}{n\rho_n^2}\left[l_{1:n}(X_{1:n}|\theta)-l_{1:n}(X_{1:n}|\theta_0)\right]\leqslant-\epsilon. \tag{2.93}$$

Proof of Proposition 2.21. This proposition is to be compared to the regularity condition imposed in Ghosh et al. (2006) (see their condition (A4) in Chapter 4). The aim of this proposition is to show that our model satisfies to a somewhat stronger version of that condition.

Let $0 < \delta$. Notice first that, similarly to what was done in (2.42), we are able to deduce that a.s.

$$\frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\widehat{\theta}_n)] \leqslant \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] =: i_n(\theta). \tag{2.94}$$

where i_n is defined over $\mathbb{R} \times [\underline{u}, \overline{u}] \times \mathbb{R}_+^* \supset \Theta \supset B^c(\theta_0, \delta\rho_n)$ by

$$i_n(\theta) = \log\frac{\sigma_0^2}{\sigma^2} + 1 + \frac{1}{n\sigma_0^2}\sum_{i=1}^{n}(\xi_i^2 - \sigma_0^2) - \frac{1}{n\sigma^2}\sum_{i=1}^{n}[\xi_i + \mu(\eta_0, t_i) - \mu(\eta, t_i)]^2. \tag{2.95}$$

$$= \log\frac{\sigma_0^2}{\sigma^2} + 1 - \frac{\sigma_0^2}{\sigma^2} + \frac{1}{n\sigma_0^2}\sum_{i=1}^{n}(\xi_i^2 - \sigma_0^2) - \frac{1}{n\sigma^2}\sum_{i=1}^{n}\left\{[\xi_i + \mu(\eta_0, t_i) - \mu(\eta, t_i)]^2 - \sigma_0^2\right\}. \tag{2.96}$$

From (2.94) it is clear that we need only prove (2.93) to end the proof.

The rest of this proof is divided into 6 major steps. Step 1 shows that for a given n the supremum considered is reached on a point θ_n. Step 2 and 3 focus on obtaining useful majorations of the supremum. Step 4 is dedicated to proving that the sequence θ_n admits an accumulation point (the coordinates of which satisfy to some conditions), while step 5 makes use of this last fact to effectively dominate the supremum. Step 6 wraps up the proof.

Step 1. We first show that a.s. for any n there exists $\theta_n \in \mathbb{R} \times [\underline{u}, \overline{u}] \times \mathbb{R}_+^*$ such that $\|\theta_n - \theta_0\| \geqslant \delta\rho_n$ and

$$i_n(\theta_n) = \sup_{\Theta \in B^c(\theta_0, \delta\rho_n)} i_n(\theta). \tag{2.97}$$

Let $n \in \mathbb{N}$ and let $(\theta_{n,k})_{k\in\mathbb{N}}$ be a sequence of points in $B^c(\theta_0, \delta\rho_n)$ such that

$$\lim_{k\longrightarrow+\infty} i_n(\theta_{n,k}) = \sup_{\Theta \in B^c(\theta_0, \delta\rho_n)} i_n(\theta).$$

From (2.95) it is obvious that $\sigma_{n,k}^2$ is bounded: if it was not, we would be able to extract a subsequence such that σ_{n,k_j}^2 would go to $+\infty$ and thus $i_n(\theta_{n,k_j})$ would go to $-\infty$. For the very same reason, $\gamma_{n,k}$ too is bounded. Recalling that $u_{n,k}$ is bounded too by definition, we now see that there exists a subsequence $(\theta_{n,k_j})_{j\in\mathbb{N}}$ in $B^c(\theta_0, \delta\rho_n)$ and a point θ_n in $\overline{B^c(\theta_0, \delta\rho_n)}$ (i.e. in $\mathbb{R} \times [\underline{u}, \overline{u}] \times \mathbb{R}_+$, and such that $\|\theta_n - \theta_0\| \geqslant \delta\rho_n$) such that $(\theta_{n,k_j})_{j\in\mathbb{N}} \xrightarrow[j\longrightarrow+\infty]{} \theta_n$.

Finally from (2.95) again it is easy to see that $\sigma_n^2 > 0$ for if it was not $i_n(\theta_{n,k_j})$ would go to $-\infty$ once again, unless (by continuity of μ with regard to η) $\xi_i + \mu(\eta_0, t_i) - \mu(\eta_n, t_i) = 0$ for all $i \leqslant n$ which a.s. does not happen.

Step 2. From the previous step and the continuity of i_n with regard to θ we are able to write

$$\sup_{\Theta \in B^c(\theta_0, \delta\rho_n)} \frac{2}{n}[l_{1:n}(X_{1:n}|\theta) - l_{1:n}(X_{1:n}|\theta_0)] = i_n(\theta_n). \tag{2.98}$$

where $(\theta_n)_{n\in\mathbb{N}}$ is the sequence defined in Step 1. We now derive a convenient majoration of $i_n(\theta_n)$. Expanding from (2.96) we get

$$\begin{aligned} i_n(\theta_n) &= \left(\log\frac{\sigma_0^2}{\sigma_n^2}+1-\frac{\sigma_0^2}{\sigma_n^2}\right)+\frac{1}{n\sigma_0^2}\sum_{i=1}^n(\xi_i^2-\sigma_0^2)-\frac{1}{n\sigma^2}\sum_{i=1}^n\left\{[\xi_i+\mu(\eta_0,t_i)-\mu(\eta,t_i)]^2-\sigma_0^2\right\}\\ &= \left(\log\frac{\sigma_0^2}{\sigma_n^2}+1-\frac{\sigma_0^2}{\sigma_n^2}\right)+\frac{\sigma_n^2-\sigma_0^2}{n\sigma_0^2\sigma_n^2}\sum_{i=1}^n(\xi_i^2-\sigma_0^2)-\frac{1}{n\sigma_n^2}\sum_{i=1}^n[\mu(\eta_0,t_i)-\mu(\eta_n,t_i)]^2\\ &\quad -\frac{2}{n\sigma_n^2}\sum_{i=1}^n[\mu(\eta_0,t_i)-\mu(\eta_n,t_i)]\xi_i \end{aligned}$$

Thanks to Lemma 2.18, we know that there exists $C_1\in\mathbb{R}_+^*$ such that

$$\begin{aligned} i_n(\theta_n) \leqslant &\left(\log\frac{\sigma_0^2}{\sigma_n^2}+1-\frac{\sigma_0^2}{\sigma_n^2}\right)+\frac{\sigma_n^2-\sigma_0^2}{n\sigma_0^2\sigma_n^2}\sum_{i=1}^n(\xi_i^2-\sigma_0^2)\\ &-\frac{1}{\sigma_n^2}C_1\|\eta_n-\eta_0\|^2-\frac{2}{n\sigma_n^2}\sum_{i=1}^n[\mu(\eta_0,t_i)-\mu(\eta_n,t_i)]\xi_i \end{aligned}$$

From there, the Law of the Iterated Logarithm and a factorisation of the last term together with Corollary 2.17 lead to:

$$\begin{aligned} i_n(\theta_n) \leqslant &\left(\log\frac{\sigma_0^2}{\sigma_n^2}+1-\frac{\sigma_0^2}{\sigma_n^2}\right)+\frac{1}{\sigma_n^2}|\sigma_n^2-\sigma_0^2|R_{1,n}-\frac{1}{\sigma_n^2}C_1\|\eta_n-\eta_0\|^2\\ &+\frac{1}{n\sigma_n^2}\left(\sum_{i=1}^n[\mu(\eta_0,t_i)-\mu(\eta_n,t_i)]^2\right)^{\frac{1}{2}}R_{2,n} \end{aligned}$$

where a.s. $R_{1,n}=\mathrm{O}\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right)$ and $R_{2,n}=\mathrm{O}\left(\log n\right)$. Lemma 2.13 ensures there exists $C_2\in\mathbb{R}_+^*$ such that

$$i_n(\theta_n) \leqslant \left(\log\frac{\sigma_0^2}{\sigma_n^2}+1-\frac{\sigma_0^2}{\sigma_n^2}\right)+\frac{1}{\sigma_n^2}|\sigma_n^2-\sigma_0^2|R_{1,n}-\frac{1}{\sigma_n^2}C_1\|\eta_n-\eta_0\|^2+\frac{1}{n\sigma_n^2}C_2n^{\frac{1}{2}}\|\eta_n-\eta_0\|R_{2,n}$$

We thus deduce that there exists $C\in\mathbb{R}_+^*$ such that:

$$i_n(\theta_n) \leqslant \left(\log\frac{\sigma_0^2}{\sigma_n^2}+1-\frac{\sigma_0^2}{\sigma_n^2}\right)-\frac{1}{\sigma_n^2}C\|\eta_n-\eta_0\|^2+\frac{1}{\sigma_n^2}\|\theta_n-\theta_0\|R_n \tag{2.99}$$

where a.s. $R_n=\mathrm{O}\left(n^{-\frac{1}{2}}\log n\right)$. Notice in particular that, due to (2.91), $R_n=\mathrm{o}(\rho_n)$.

Step 3. We obtain two majorations, (2.101) and (2.102), that we will make use of in the coming steps. Using a conversion of $\theta=(\gamma,u,\sigma^2)$ into the spherical coordinate system we write θ_n as

$$\theta_n=(r_n\cos\psi_n\cos\phi_n,r_n\sin\psi_n\cos\phi_n,r_n\sin\phi_n),$$

where

$$(r_n,\psi_n,\phi_n)\in\mathbb{R}_+^*\times[0,2\pi]\times]0,\pi[,$$

and deduce from (2.99) that

$$i_n(\theta_n) \leqslant \left(\log \frac{\sigma_0^2}{r_n \sin\phi_n} + 1 - \frac{\sigma_0^2}{r_n \sin\phi_n}\right) - Cr_n \frac{\cos^2\phi_n}{\sin\phi_n} + \frac{1}{\sin\phi_n} R_n \tag{2.100}$$

$$\leqslant \left(\log \frac{\sigma_0^2}{r_n \sin\phi_n} + 1 - \frac{\sigma_0^2}{r_n \sin\phi_n}\right) + \frac{1}{\sin\phi_n}\left[R_n - Cr_n\cos^2\phi_n\right]. \tag{2.101}$$

From (2.100) we also get the following majoration

$$i_n(\theta_n) \leqslant \left(\log \frac{\sigma_0^2}{r_n \sin\phi_n} + 1 - \frac{\sigma_0^2}{r_n \sin\phi_n}\right) + \frac{1}{\sin\phi_n} R_n. \tag{2.102}$$

Step 4. We show that the sequence $(\theta_n)_{n\in\mathbb{N}}$ we built, converges to a finite limit θ_∞ (extracting a subsequence if necessary). Extracting a subsequence if necessary, we can assume that $(\psi_n, \phi_n) \to (\psi_\infty, \phi_\infty) \in [0, 2\pi] \times [0, \pi]$. We consider the two following mutually exclusive situations.

Situation A: $\phi_\infty = 0 \mod \pi$. In this situation, there exists $\epsilon > 0$ such that for any n large enough,

$$\begin{aligned}\left[R_n - Cr_n\cos^2\phi_n\right] &= \left(\frac{R_n}{r_n} - C\cos^2\phi_n\right) r_n \\ &\leqslant -\epsilon r_n,\end{aligned}$$

because a.s. $R_n = o(r_n)$ (since $R_n = o(\rho_n)$ and $r_n \leqslant \rho_n$). Used together with (2.101), this leads to

$$i_n(\theta_n) \leqslant \left(\frac{\sigma_0^2}{r_n \sin\phi_n} - 1 - \log \frac{\sigma_0^2}{r_n \sin\phi_n}\right) - \epsilon \frac{r_n}{\sin\phi_n},$$

for any n large enough and hence $i_n(\theta_n) \to -\infty$ whether r_n goes to zero or not.

Situation B: $\phi_\infty \neq 0 \mod \pi$. In this situation, from (2.102), we see that $r_n \to 0$ and $r_n \to +\infty$ both lead to $i_n(\theta_n) \to -\infty$.

Observing that $i_n(\theta)$ converges a.s. to a finite value for any $\theta \in \Theta$ as $n \to +\infty$, we see that $\lim_{n\to\infty} i_n(\theta_n) = -\infty$ is not possible by construction of the sequence θ_n, and deduce that, extracting a subsequence if necessary, there exists

$$(r_\infty, \psi_\infty, \phi_\infty) \in \mathbb{R}_+^* \times [0, 2\pi] \times]0, \pi[,$$

such that $\lim_{n\to+\infty} \theta_n = \theta_\infty$. Notice that in particular, $\sigma_\infty^2 > 0$.

Step 5. We will now end the proof by showing that there exists $\epsilon > 0$ such that for any n large enough

$$i_n(\theta_n) \leqslant -\epsilon\rho_n^2. \tag{2.103}$$

We consider the two following mutually exclusive situations.

Situation A: $\sigma_\infty^2 \neq \sigma_0^2$. In this situation, from (2.99) we get

$$i_n(\theta_n) \leqslant \left(\log \frac{\sigma_0^2}{\sigma_n^2} + 1 - \frac{\sigma_0^2}{\sigma_n^2}\right) + \frac{1}{\sigma_n^2} \|\theta_n - \theta_0\| R_n$$

and the right-hand side converges to

$$\left(\log \frac{\sigma_0^2}{\sigma_\infty^2} + 1 - \frac{\sigma_0^2}{\sigma_\infty^2}\right) < 0.$$

There hence exists $\epsilon > 0$ such that for any n large enough

$$i_n(\theta_n) \leqslant -\epsilon.$$

Since $\rho_n = \mathrm{O}(1)$ by (2.90), (2.103) is a direct consequence of this.

Situation B: $\sigma_\infty^2 = \sigma_0^2$. In this situation, recalling that for any $x > 0$

$$\log x + 1 - x \leqslant -\frac{(x-1)^2}{2} + \frac{(x-1)^3}{3},$$

we deduce from (2.99) that for any n large enough

$$\begin{aligned}
i_n(\theta_n) &\leqslant -\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_n^2} - 1\right)^2 + \frac{1}{3}\left(\frac{\sigma_0^2}{\sigma_n^2} - 1\right)^3 - \frac{1}{\sigma_n^2} C\|\eta_n - \eta_0\|^2 + \frac{1}{\sigma_n^2} \|\theta_n - \theta_0\| R_n \\
&\leqslant \left(\frac{\sigma_0^2}{\sigma_n^2} - 1\right)^2 \left[\frac{1}{3}\left(\frac{\sigma_0^2}{\sigma_n^2} - 1\right) - \frac{1}{2}\right] - \frac{1}{\sigma_n^2} C\|\eta_n - \eta_0\|^2 + \frac{1}{\sigma_n^2} \|\theta_n - \theta_0\| R_n \\
&\leqslant -\frac{1}{4}\left(\frac{\sigma_0^2}{\sigma_n^2} - 1\right)^2 - \frac{1}{\sigma_n^2} C\|\eta_n - \eta_0\|^2 + \frac{1}{\sigma_n^2} \|\theta_n - \theta_0\| R_n \\
&\leqslant \frac{1}{\sigma_n^2}\left\{-c\left[(\sigma_0^2 - \sigma_n^2)^2 - \|\eta_n - \eta_0\|^2\right] + \|\theta_n - \theta_0\| R_n\right\}
\end{aligned}$$

where $c = \min(1/4, C) > 0$. It follows that for any n large enough

$$\begin{aligned}
i_n(\theta_n) &\leqslant \frac{1}{\sigma_n^2}\left(-c\|\theta_n - \theta_0\|^2 + \|\theta_n - \theta_0\| R_n\right) \\
&\leqslant \frac{1}{\sigma_n^2} \|\theta_n - \theta_0\| \left(R_n - c\|\theta_n - \theta_0\|\right).
\end{aligned}$$

Thus, for any n large enough

$$i_n(\theta_n) \leqslant \frac{1}{\sigma_n^2} \frac{\|\theta_n - \theta_0\|}{\rho_n} \left(\frac{R_n}{\rho_n} - c\frac{\|\theta_n - \theta_0\|}{\rho_n}\right) \rho_n^2.$$

Recalling that

$$\begin{aligned}
\|\theta_n - \theta_0\| &\geqslant \delta\rho_n, \\
R_n &= \mathrm{o}(\rho_n), \\
\sigma_n^2 &\to \sigma_\infty^2 > 0
\end{aligned}$$

we obtain for any n large enough,

$$i_n(\theta_n) \leqslant \frac{1}{\sigma_n^2} \frac{\|\theta_n - \theta_0\|}{\rho_n} \left(-c\frac{\delta}{2}\right) \rho_n^2 \leqslant -\frac{c\delta^2}{2\sigma_n^2}\rho_n^2 \leqslant -\frac{c\delta^2}{3\sigma_\infty^2}\rho_n^2.$$

Hence (2.103) holds in this situation too: it suffices to take $\epsilon = \dfrac{c\delta^2}{3\sigma_\infty^2}$.

We just proved that (2.103) holds in both cases considered.

Step 6. (2.93) is a consequence of (2.98) and (2.103).

■

Lemma 2.22. *Let* $0 < \delta < 1$ *then under Assumptions A1–A4 and conditions* (2.10)*, a.s. there exists a constant* $C \in R_+^*$ *such that for any n large enough and for any* $1 \leqslant i_1, i_2, i_3 \leqslant 3$

$$\left| \frac{1}{n} \frac{\partial^3 l_{1:n}^*(X_{1:n}|\theta)}{\partial_{i_1}\partial_{i_2}\partial_{i_3}} \right| \leqslant C \tag{2.104}$$

for any $\theta \in B(\theta_0, \delta d_n)$.

Proof of Lemma 2.22. Let $0 < \delta < 1$. We will prove (2.104) stands true for any $1 \leqslant i_1, i_2, i_3 \leqslant 3$. First notice that for n large enough, $\theta \mapsto l_{1:n}^*(X_{1:n}|\theta)$ is indeed infinitely continuously differentiable over $B(\theta_0, \delta d_n)$ by definition of the pseudo-problem. Any θ subsequently considered within this proof is assumed to belong to $B(\theta_0, \delta d_n)$. Any convergence subsequently mentioned within this proof is uniform in θ for $\theta \in B(\theta_0, \delta d_n)$ for any n large enough thanks to Theorem 2.11 and Lemma 2.16.

Proof of (2.104) *for* $\beta = (3, 0, 0)$.

$$\frac{1}{n} \frac{\partial^3 l_{1:n}^*(X_{1:n}|\theta)}{(\partial\gamma)^3} = 0.$$

Proof of (2.104) *for* $\beta = (2, 1, 0)$.

$$\left| \frac{1}{n} \frac{\partial^3 l_{1:n}^*(X_{1:n}|\theta)}{(\partial\gamma)^2 \partial u} \right| = \frac{2}{\sigma^2} \left| \frac{1}{n} \sum_{i=1}^{n^*} (t_i - u) \mathbb{1}_{]t_i, +\infty[}(u) \right| \xrightarrow[n \to +\infty]{} \frac{2}{\sigma^4} \left| \int_{\underline{u}}^{u} (t - u) f(t) \, \mathrm{d}t \right| \leqslant \frac{2}{\sigma^2} |\overline{u} - \underline{u}|.$$

Proof of (2.104) *for* $\beta = (2, 0, 1)$.

$$\left| \frac{1}{n} \frac{\partial^3 l_{1:n}^*(X_{1:n}|\theta)}{(\partial\gamma)^2 \partial\sigma^2} \right| = \frac{1}{\sigma^4} \left| \frac{1}{n} \sum_{i=1}^{n^*} (t_i - u)^2 \mathbb{1}_{]t_i, +\infty[}(u) \right| \xrightarrow[n \to +\infty]{} \frac{1}{\sigma^4} \left| \int_{\underline{u}}^{u} (t - u)^2 f(t) \, \mathrm{d}t \right| \leqslant \frac{1}{\sigma^4} |\overline{u} - \underline{u}|^2.$$

Proof of (2.104) *for* $\beta = (1, 2, 0)$.

$$\left| \frac{1}{n} \frac{\partial^3 l_{1:n}^*(X_{1:n}|\theta)}{\partial\gamma \cdot (\partial u)^2} \right| = \frac{2}{\sigma^2} \left| \gamma \frac{1}{n} \sum_{i=1}^{n^*} \mathbb{1}_{]t_i, +\infty[}(u) \right| \xrightarrow[n \to +\infty]{} \frac{2}{\sigma^2} \left| \gamma \int_{\underline{u}}^{u} f(t) \, \mathrm{d}t \right| \leqslant \frac{2}{\sigma^2} |\gamma|.$$

Proof of (2.104) *for* $\beta = (1,1,1)$.

$$\begin{aligned}\left|\frac{1}{n}\frac{\partial^3 l^*_{1:n}(X_{1:n}|\theta)}{\partial\gamma\partial u\partial\sigma^2}\right| &= \frac{1}{\sigma^4}\left|\frac{1}{n}\sum_{i=1}^{n^*}(X_i - 2\gamma\cdot(t_i-u))\mathbb{1}_{]t_i,+\infty[}(u)\right|\\ &= \frac{1}{\sigma^4}\left|\frac{1}{n}\sum_{i=1}^{n^*}\left[\xi_i + \gamma_0\cdot(t_i-u_0)\mathbb{1}_{]t_i,+\infty[}(u_0) - 2\gamma\cdot(t_i-u)\right]\mathbb{1}_{]t_i,+\infty[}(u)\right|\\ &\xrightarrow[n\to+\infty]{a.s} \frac{1}{\sigma^4}\left|\int_{\underline{u}}^{\min(u,u_0)}\gamma_0\cdot(t-u_0)f(t)\,\mathrm{d}t - 2\int_{\underline{u}}^{u}\gamma\cdot(t-u)f(t)\,\mathrm{d}t\right|\end{aligned}$$

And this limit is bounded by $\frac{3}{\sigma^4}|\overline{u}-\underline{u}|(|\gamma|+|\gamma_0|)$.

Proof of (2.104) *for* $\beta = (1,0,2)$.

$$\begin{aligned}\left|\frac{1}{n}\frac{\partial^3 l^*_{1:n}(X_{1:n}|\theta)}{\partial\gamma\cdot(\partial\sigma^2)^2}\right| &= \frac{2}{\sigma^6}\left|\frac{1}{n}\sum_{i=1}^{n^*}\left[x_i - \gamma\cdot(t_i-u)\mathbb{1}_{]t_i,+\infty[}(u)\right](t_i-u)\mathbb{1}_{]t_i,+\infty[}(u)\right|\\ &= \frac{2}{\sigma^6}\left|\frac{1}{n}\sum_{i=1}^{n^*}\left[\xi_i + \gamma_0\cdot(t_i-u_0)\mathbb{1}_{]t_i,+\infty[}(u_0) - \gamma\cdot(t_i-u)\right](t_i-u)\mathbb{1}_{]t_i,+\infty[}(u)\right|\\ &\xrightarrow[n\to+\infty]{a.s} \frac{2}{\sigma^6}\left|\int_{\underline{u}}^{\min(u,u_0)}\gamma_0\cdot(t-u_0)(t-u)f(t)\,\mathrm{d}t - \int_{\underline{u}}^{u}\gamma\cdot(t-u)^2f(t)\,\mathrm{d}t\right|\end{aligned}$$

And this limit is bounded by $\frac{4}{\sigma^6}|\overline{u}-\underline{u}|^2(|\gamma|+|\gamma_0|)$.

Proof of (2.104) *for* $\beta = (0,3,0)$.

$$\left|\frac{1}{n}\frac{\partial^3 l^*_{1:n}(X_{1:n}|\theta)}{(\partial u)^3}\right| = 0.$$

Proof of (2.104) *for* $\beta = (0,2,1)$.

$$\left|\frac{1}{n}\frac{\partial^3 l^*_{1:n}(X_{1:n}|\theta)}{(\partial u)^2\partial\sigma^2}\right| = \frac{1}{\sigma^4}\gamma^2\left|\frac{1}{n}\sum_{i=1}^{n^*}\mathbb{1}_{]t_i,+\infty[}(u)\right| \xrightarrow[n\to+\infty]{} \frac{1}{\sigma^4}\gamma^2\left|\int_{\underline{u}}^{u}f(t)\,\mathrm{d}t\right| \leqslant \frac{1}{\sigma^4}\gamma^2.$$

Proof of (2.104) *for* $\beta = (0,1,2)$.

$$\begin{aligned}\left|\frac{1}{n}\frac{\partial^3 l^*_{1:n}(X_{1:n}|\theta)}{\partial u(\partial\sigma^2)^2}\right| &= \frac{2}{\sigma^6}\left|\frac{1}{n}\sum_{i=1}^{n^*}\left[x_i - \gamma\cdot(t_i-u)\mathbb{1}_{]t_i,+\infty[}(u)\right]\gamma\mathbb{1}_{]t_i,+\infty[}(u)\right|\\ &= \frac{2}{\sigma^6}\left|\frac{1}{n}\sum_{i=1}^{n^*}\left[\xi_i + \gamma_0\cdot(t_i-u_0)\mathbb{1}_{]t_i,+\infty[}(u_0) - \gamma\cdot(t_i-u)\right]\gamma\mathbb{1}_{]t_i,+\infty[}(u)\right|\\ &\xrightarrow[n\to+\infty]{a.s} \frac{2}{\sigma^6}\left|\int_{\underline{u}}^{u_0}\gamma\gamma_0\cdot(t-u_0)f(t)\,\mathrm{d}t - \int_{\underline{u}}^{u}\gamma^2(t-u)f(t)\,\mathrm{d}t\right|\end{aligned}$$

And this limit is bounded by $\frac{2}{\sigma^6}|\overline{u}-\underline{u}|(|\gamma^2|+|\gamma_0\gamma|)$.

Proof of (2.104) *for* $\beta = (0,0,3)$.

$$\begin{aligned}\left|\frac{1}{n}\frac{\partial^3 l^*_{1:n}(X_{1:n}|\theta)}{(\partial\sigma^2)^3}\right| &= \frac{1}{\sigma^6}\left|-1+\frac{3}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n^*}\left[x_i-\gamma\cdot(t_i-u)\mathbb{1}_{]t_i,+\infty[}(u)\right]^2\right| \\ &= \frac{1}{\sigma^6}\left|-1+\frac{3}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n^*}\left[\xi_i+\gamma\cdot(t_i-u_0)\mathbb{1}_{]t_i,+\infty[}(u_0)-\gamma\cdot(t_i-u)\mathbb{1}_{]t_i,+\infty[}(u)\right]^2\right| \\ &\xrightarrow[n\to+\infty]{a.s} \frac{1}{\sigma^8}\left|-\sigma^2+3\left(\sigma_0^2+\int_{\underline{u}}^{u_0}\gamma_0^2(t-u_0)^2f(t)\,\mathrm{d}t\right.\right. \\ &\qquad\left.\left.-2\int_{\underline{u}}^{\min(u,u_0)}\gamma\gamma_0\cdot(t-u_0)(t-u)f(t)\,\mathrm{d}t+\int_{\underline{u}}^{u}\gamma^2(t-u)^2f(t)\,\mathrm{d}t\right)\right|\end{aligned}$$

And this limit is bounded by $\frac{1}{\sigma^8}\left[3\sigma_0^2+\sigma^2+(|\gamma|+|\gamma_0|)^2(\overline{u}-\underline{u})^2\right]$.

(2.104) is thus a direct consequence of both the uniform convergences mentioned above and the trivial majoration of all the limits involved by a fixed constant C for any n large enough. ■

Acknowledgements

The authors would like to thank the referees and associate editor for their constructive comments.

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

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.

3.1 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); 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 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 Bruhns et al. (2005); Menage et al. (1988). Other weather-related models include the works of Engle et al. (1986) 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 3.2 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 3.3, 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 3.4, 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 3.5, 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.

3.2 METHODOLOGY

3.2.1 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 3.1 (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 3.2 (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*

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

3.2.2 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

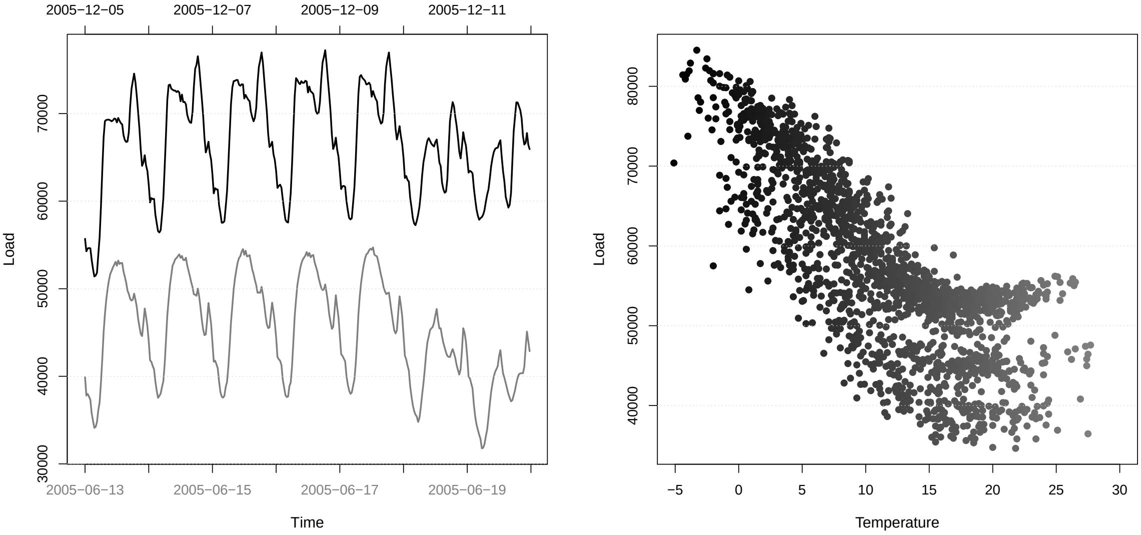


Figure 3.1: Left : French Electricity load from 13/06/2005 to 29/06/2005 (in grey) and from 05/12/2005 to 11/12/2005 (in black). The load is expressed in MW. Notice the daily patterns of the electricity load are not the same during summer and winter. Right : French Electricity load at 10:00 over 5 years against temperatures. The load seems to increase linearly with the temperature below a certain threshold.

follows: for $t = 1, \dots, N$,

$$\begin{aligned} y_t &= x_t^{(1)} x_t^{(2)} + x_t^{(3)} + \epsilon_t \\ x_t^{(1)} &= \sum_{j=1}^{d_{11}} \left[z_j^{\cos} \cos\left(\frac{2j\pi}{365.25}t\right) + z_j^{\sin} \sin\left(\frac{2j\pi}{365.25}t\right) \right] + \sum_{j=1}^{d_{12}} \omega_j \mathbb{1}_{\Omega_j}(t), \\ x_t^{(2)} &= \sum_{j=1}^{d_2} \psi_j \mathbb{1}_{\Psi_j}(t), \\ x_t^{(3)} &= g(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u), \end{aligned} \tag{3.1}$$

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 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 3.1 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 3.1 too. For obvious identifiability reasons, the vector ψ is restricted to the positive quadrant of the $\| \cdot \|_1$-unit sphere in $\mathbb{R}^{d_2}$, that we denote

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

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

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

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 3.1.

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 $t = 1, \ldots, N$,

$$y_t = (A_{t\bullet}\alpha)(B_{t\bullet}\beta + C_t) + \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u) + \epsilon_t. \tag{3.2}$$

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 $B_{+}^{d_\beta}(0,1) = \{\beta \in (\mathbb{R}_{+})^{d_\beta};\ \|\beta\|_1 \leqslant 1\}$ 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 3.1), 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 3.5 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 (3.2), 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

$$y_t = f_t(\eta) + \epsilon_t \tag{3.3}$$

and use the notation $f(\eta) = (f_1(\eta), \dots, f_N(\eta))$ 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 (3.2) reads:

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

3.3 SPECIFICATIONS OF THE PRIORS

3.3.1 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 (3.2), and then prove that it leads to a proper posterior distribution (see Proposition 3.3). Following the methodology exposed in Section 3.2.1, 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

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

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:

$$l \sim \mathcal{G}(a_l, b_l), \qquad q \sim \mathcal{N}(1, \sigma_q^2), \qquad r \sim \mathcal{G}(a_r, b_r), \tag{3.4}$$

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:

$$\pi(\theta, k, l, q, r) \propto \pi(\eta|k,l)\pi(k|q,r)\pi(l)\pi(q)\pi(r)\pi(\sigma^2) \tag{3.5}$$

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 3.2, the posterior measure is given by

$$
\begin{aligned}
\pi(\theta,k,l,q,r|y) \propto\ & L(y|\theta)\pi(\theta,k,l,q,r) \qquad (3.6)\\
\propto\ & \sigma^{-N-2} \exp\left(-\frac{1}{2}\sigma^{-2}\|y - f(\eta)\|_2^2\right) \mathbb{1}_{[0,1]\times[\underline{u},\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1, u, \sigma^2)\\
& \times l^{\frac{d}{2}} \exp\left(-\frac{1}{2}(\eta - K\mu^{\mathcal{A}})' l (\Sigma^{\mathcal{A}})^{-1}(\eta - K\mu^{\mathcal{A}})\right)\\
& \times |r|^{\frac{d}{2}} \exp\left(-\frac{1}{2} r \sum_{i=1}^{d}(k_i - q)^2\right) l^{a_l - 1}\exp\left(-b_l l\right)\mathbb{1}_{\mathbb{R}_+^*}(l)\\
& \times |\sigma_q^{-2}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sigma_q^{-2}(q-1)^2\right) r^{a_r-1}\exp\left(-b_r r\right)\mathbb{1}_{\mathbb{R}_+^*}(r).
\end{aligned}
$$

Proposition 3.3. *For $(\beta, u) \in B_+^{d_\beta}(0,1) \times [\underline{u},\ \overline{u}]$ denote $A_*(\beta,u)$ the matrix whose rows are*

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

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 (3.2) *and the posterior measure* (3.6) *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. ∎

3.3.2 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 asymptotic 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, Launay et al. (2012a) show that this issue can be dealt with and prove 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 3.4).

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

$$\begin{aligned}\pi(\theta|y) &\propto L(y|\theta)\pi(\theta) \\ &\propto \sigma^{-N-2}\exp\left(-\frac{1}{2}\sigma^{-2}\|y-f(\eta)\|_2^2\right)\mathbb{1}_{[0,1]\times[\underline{u},\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1,u,\sigma^2).\end{aligned} \tag{3.7}$$

Proposition 3.4. *For $(\beta,u)\in B_+^{d_\beta}(0,1)\times[\underline{u},\,\overline{u}]$ denote $A_*(\beta,u)$ the matrix whose rows are*

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

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 (3.2)*, the posterior measure* (3.7) *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

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

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

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

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

$$\|y-f(\eta)\|_2^{-N}\sim\|y-A_*(\beta_0,u_0)\alpha_*\|_2^{-N}. \tag{3.8}$$

The triangular inequality applied to the right hand side of (3.8) gives

$$\|y - A_*(\beta_0, u_0)\alpha_*\|_2^{-N} \leqslant \big| \|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2 \big|^{-N}. \tag{3.9}$$

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

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

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

$$\big| \|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2 \big|^{-N} \sim \lambda^{-N/2} \|\alpha_*\|_2^{-N}. \tag{3.10}$$

Combining (3.8), (3.9) and (3.10) together, we see that the integrability of the left hand side of (3.8) 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. 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.

3.4 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 (3.2) 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.

3.4.1 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 $y = (y_1, \dots, y_N)$, 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:

$$\begin{aligned} \widehat{y}_{N+k} &:= \mathbb{E}^{\pi}[y_{N+k}|y] && (3.11)\\ &= \int y_{N+k}\,g(y_{N+k}|y)\,\mathrm{d}y_{N+k}. && (3.12) \end{aligned}$$

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.:

$$\sqrt{\frac{1}{365}\sum_{i=1}^{365}[f_{N+i}(\eta_0) - \widehat{y}_{N+i}]^2}. \tag{3.13}$$

3.4.2 Construction of simulated datasets

Both datasets $\mathcal{A}$ and $\mathcal{B}$ were simulated according to the model (3.2) 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 (3.2).

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:

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

where the coordinates λ 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.

3.4.3 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 3.2 shows the posterior error of η (posterior mean minus the true value) of η, based on 300 replications that correspond to the case where $\lambda_\alpha = \lambda_\beta = \lambda_\gamma = \lambda_u = 1$ 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.

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 3.3 shows

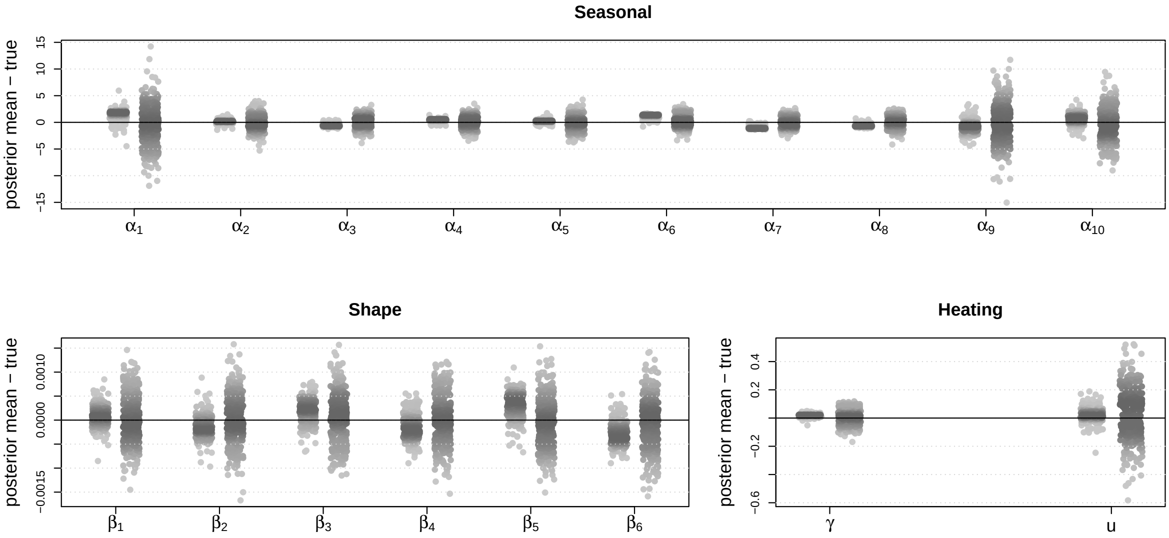


Figure 3.2: The posterior error (posterior mean minus true value) of α (seasonal parameters), β (shape parameters), and γ and u (heating parameters), based on 300 replications. Leftmost replications correspond to the informative method while the rightmost replications correspond to the non informative method. Here $\lambda_\alpha = \lambda_\beta = \lambda_\gamma = \lambda_u = 1$.

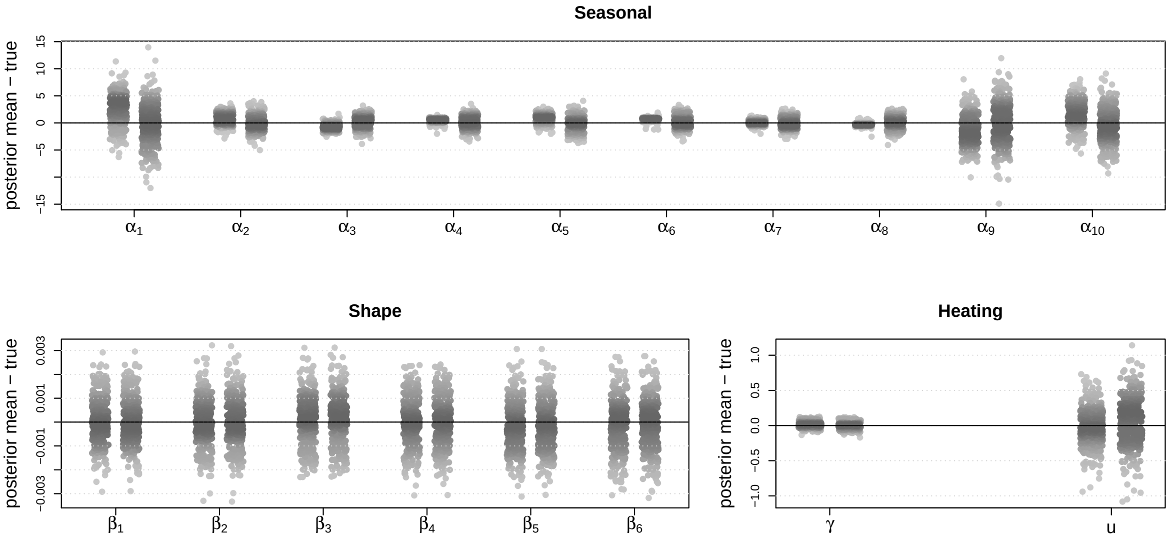


Figure 3.3: Same caption as in Figure 3.2 except $\lambda_\beta = \lambda_u = 1$ and $\lambda_\alpha = \lambda_\gamma = 0.5$.

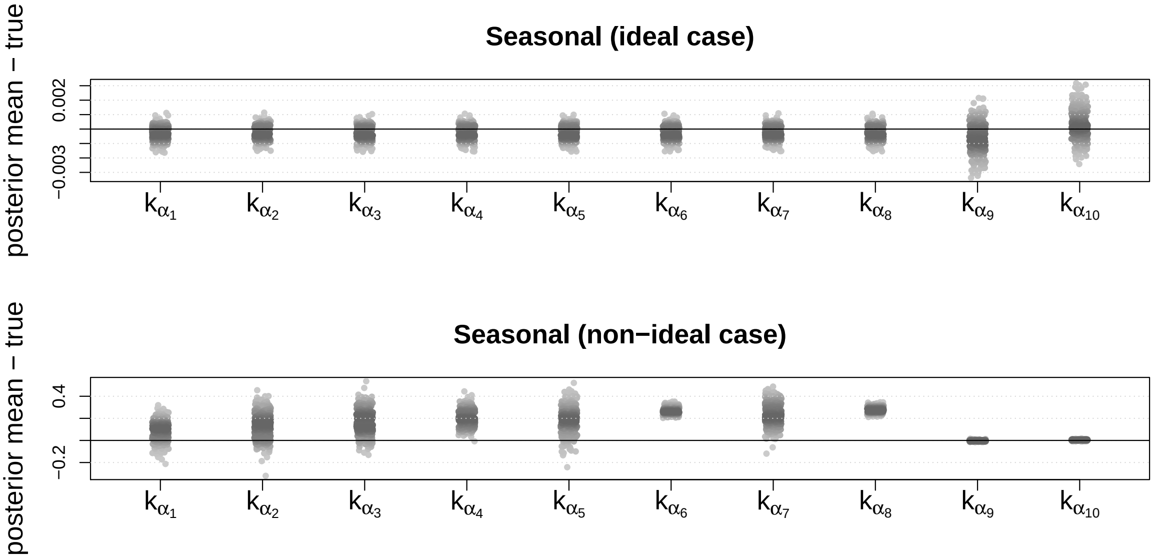


Figure 3.4: The posterior mean of coefficients k minus the true value of the corresponding λ for the seasonal parameters, based on 300 replications. Top row is for the case were $\lambda_\alpha = \lambda_\beta = \lambda_\gamma = \lambda_u = 1$ and bottom row is for the case where $\lambda_\beta = \lambda_u = 1$ and $\lambda_\alpha = \lambda_\gamma = 0.5$. Leftmost replications correspond to the informative method while the rightmost replications correspond to the non informative method. Results for the other coordinates 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).

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 $\lambda_\beta = \lambda_u = 1$ and $\lambda_\alpha = \lambda_\gamma = 0.5$. Situations such as $\lambda_\alpha = \lambda_\gamma = \lambda_u = 1$ and $\lambda_\beta = 0.5$ or $\lambda_\alpha = \lambda_\gamma = \lambda_\beta = 1$ and $\lambda_u = 0.5$ 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 3.3.1 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 $\lambda_\alpha = \lambda_\beta = \lambda_\gamma = \lambda_u = 1$ as illustrated in the top row of Figure 3.4 which shows the posterior error of k. When $\lambda_\beta = \lambda_u = 1$ and $\lambda_\alpha = \lambda_\gamma = 0.5$, the estimations obtained are of lesser quality as demonstrated in the bottom row of Figure 3.4: 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

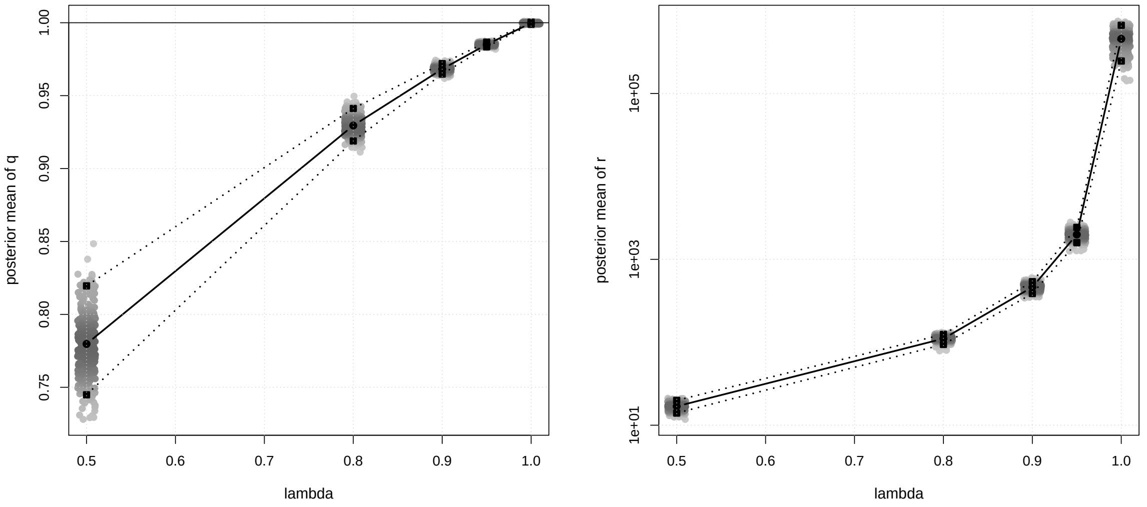


Figure 3.5: In grey: posterior mean of q (left) and r (right, on a log scale) for the informative prior (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $\lambda_\alpha = \lambda_\gamma$ tested. In black: the circles correspond to the averages, while the squares correspond to the 5% and 95% empirical quantiles. Here $\lambda_\beta = \lambda_u = 1$.

over the 300 replications for each of the five simulated scenarios and no reasonable conclusion could be drawn from it.

On the other hand, the estimation of the hyperparameter q does reveal a bit of information about the two datasets $\mathcal{A}$ and $\mathcal{B}$. It is the mean of the coordinates of k on the real axis, as can be seen in the definition of the informative prior in (3.5) on page 77. 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 3.5 shows the evolution of the posterior mean of q as $\lambda_\alpha = \lambda_u$ ranges over [0.5, 1].

The estimation of the hyperparameter r (inverse-variance of the prior distribution on k, see (3.5) 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 3.5 shows a clear decline of r when $\lambda_\alpha = \lambda_u$ 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.

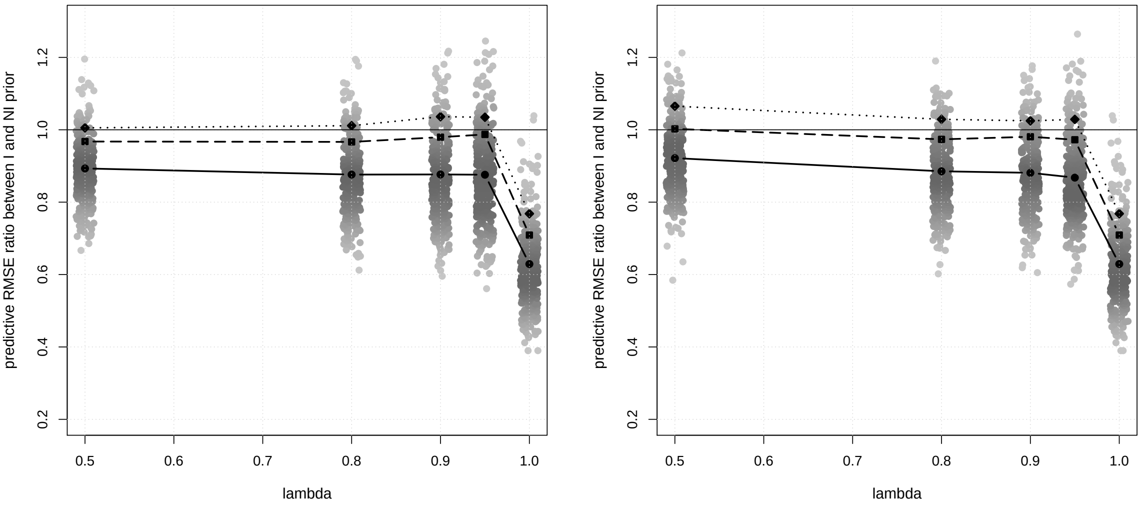


Figure 3.6: In grey: ratio between error predictions for the informative and the non informative approach (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $\lambda_\alpha = \lambda_\gamma$ (left, where $\lambda_\beta = \lambda_u = 1$) and λ_u (right, where $\lambda_\alpha = \lambda_\beta = \lambda_\gamma = 1$) tested. In black: circles correspond to the averages, while squares and diamonds correspond to the 80% and 90% empirical quantiles of these ratios.

Prediction. We compared the informative and the non informative models using our comparison criterion defined in (3.13) and computing the ratio between the two models for different values of λ_α and λ_γ, λ_β and λ_u being both set to 1. The left part of Figure 3.6 shows the results we obtained for λ_α and λ_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 of the informative hierarchical prior still leads to an average improvement of 15% over the non informative model, as can be seen on Figure 3.6. The results obtained when λ_β or λ_u are varying while the other coordinates of k are fixed to 1 were very similar (see for example the right part of Figure 3.6).

3.5 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.

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

Table 3.1: Sample size (in days) of the datasets for both experiments.

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 subset of $\mathcal{A}$ and represents around 50% of the total load of $\mathcal{A}$.

3.5.1 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 the dataset for this first application, as we wanted to benchmark our method against others on a smooth and rather easy-going signal, so as not to put any one method at a disadvantage due to the signal's specificity. 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 16°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 3.5.2), 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 3.1 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 3.3.2 (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.

It is clear from the results exposed in Table 3.2 that the informative prior outperforms all the alternative methods by a large margin in each case. Figure 3.7 shows that the Bayesian informative

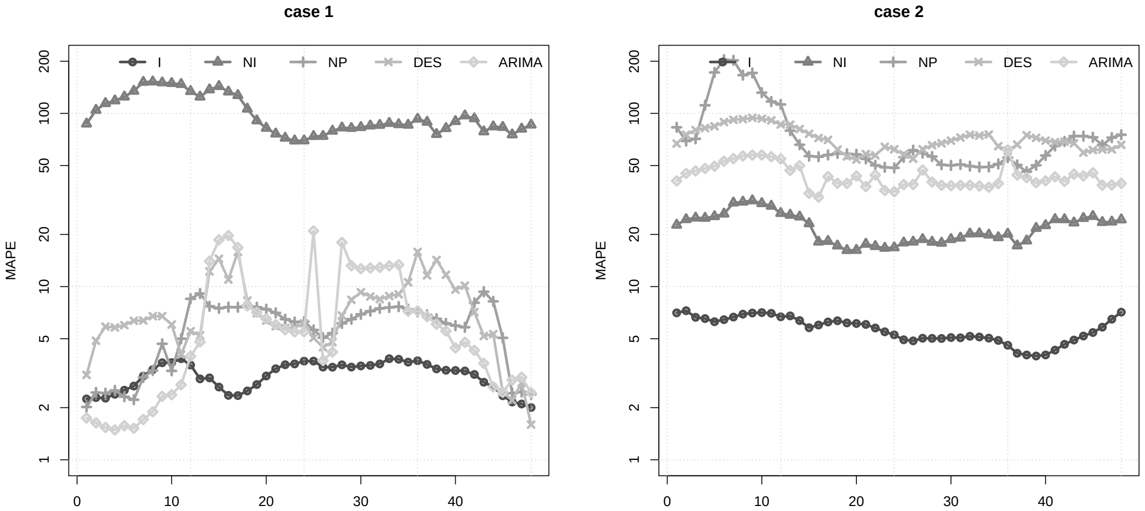


Figure 3.7: Quality of the predictions (MAPE in %) averaged for each instant, for both experiments: case 1 (left) and case 2 (right). The ordinate axis is in log-scale. Each shade of gray corresponds to one of the 5 methods tested.

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

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.

3.5.2 Role of the hyperparameters

For this application, the setup was the same as the one described at the start of Section 3.5.1. 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 3.5.1, 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 3.8 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 3.8 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 3.9 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 3.4 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.

3.5.3 Role of the sample size

For this application, the setup is almost identical to the one described at the start of Section 3.5.1 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 18°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. Note that for $\mathcal{A}$ we used the same dataset that we used for our two first applications.

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. The diagram in Figure 3.10 describes the 4 scenarios considered.

The non informative prior leads to a better fit than the informative prior as can be seen in Table 3.3. 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 3.3 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

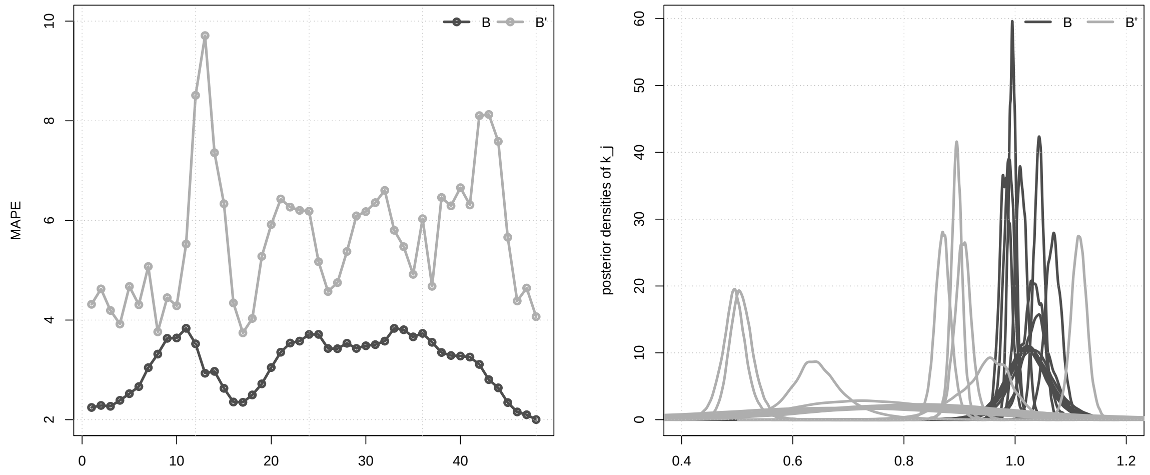


Figure 3.8: Quality of the predictions (MAPE in %) averaged for each instant, for populations $\mathcal{B}$ and $\mathcal{B}'$ (left). Posterior densities of the similarity coefficients k_j for populations $\mathcal{B}$ and $\mathcal{B}'$ at midday (right).

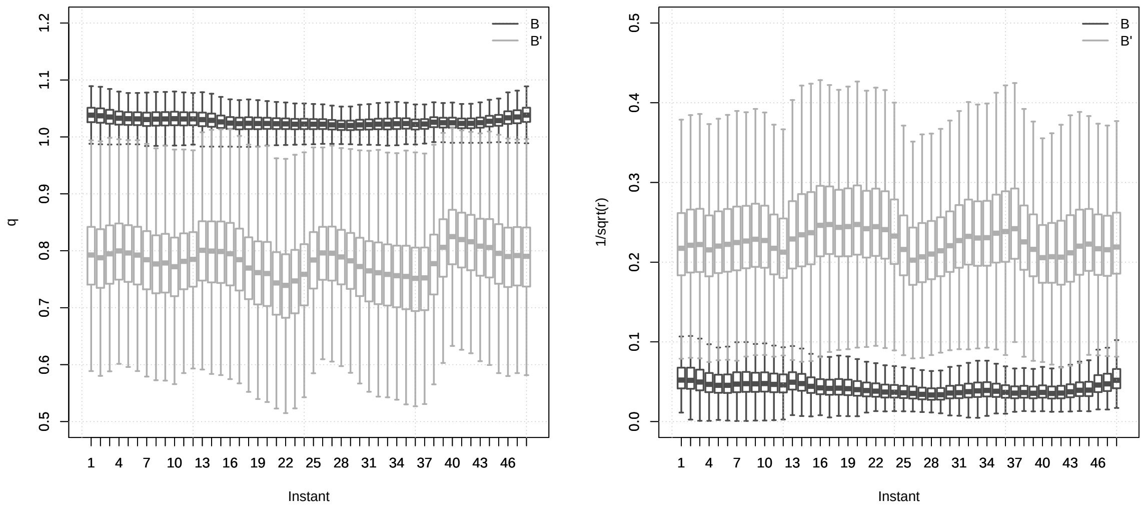


Figure 3.9: Boxplots of the posterior densities of q (left) and $1/\sqrt{r}$ (right) (mean and standard deviation of the similarity coefficients k_j) for populations $\mathcal{B}$ and $\mathcal{B}'$, throughout the 48 instants of the day.

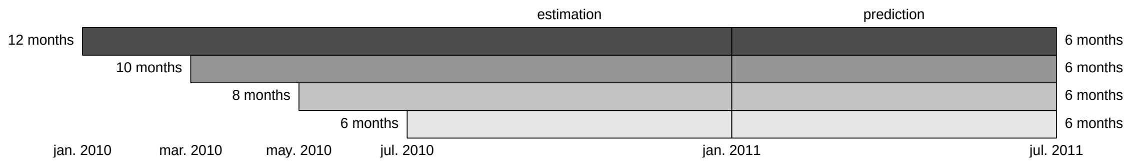


Figure 3.10: Ranges of the estimation (from 12 to 6 months) and prediction (6 months) time-windows for the 4 scenarios considered.

predictive quality for the informative prior remains moderate and stable.

	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 3.3: 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 3.13 shows the average error in prediction for each month while Figures 3.11 and 3.12 display the average error in prediction for each half hour (from 00:00 to 23:30). It is important to note that the use of the non informative prior leads to overfitting the model: the results presented in Table 3.3 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 revealed 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 model more robust with regard 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.

3.6 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.

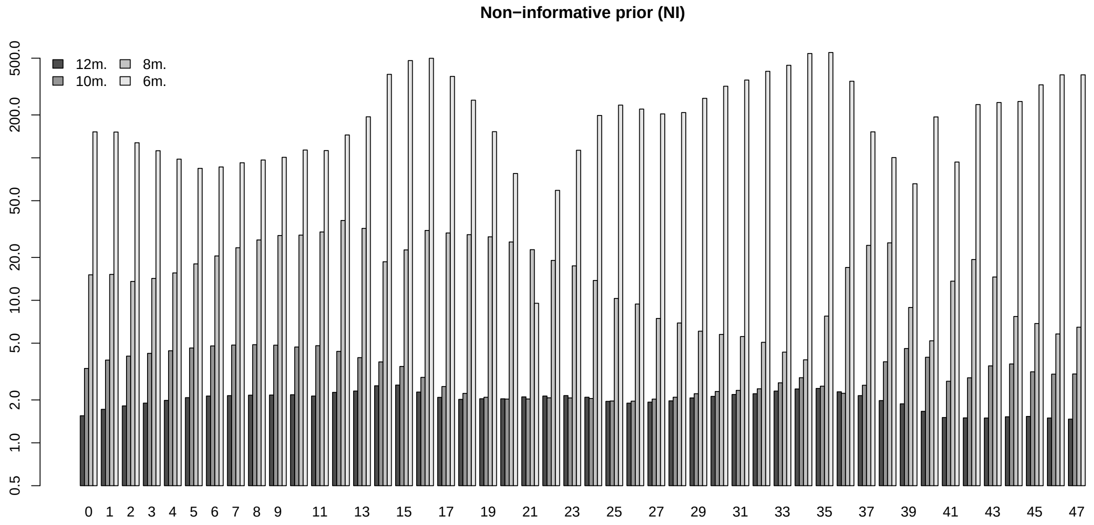


Figure 3.11: Using the non informative prior: quality of the predictions (MAPE in %) averaged for each instant (all the 180 or so days within the prediction time-window are used for those averages), with an estimation period ranging from 12 to 6 months. The ordinate axis is in log-scale. Each shade of gray corresponds to a different scenario.

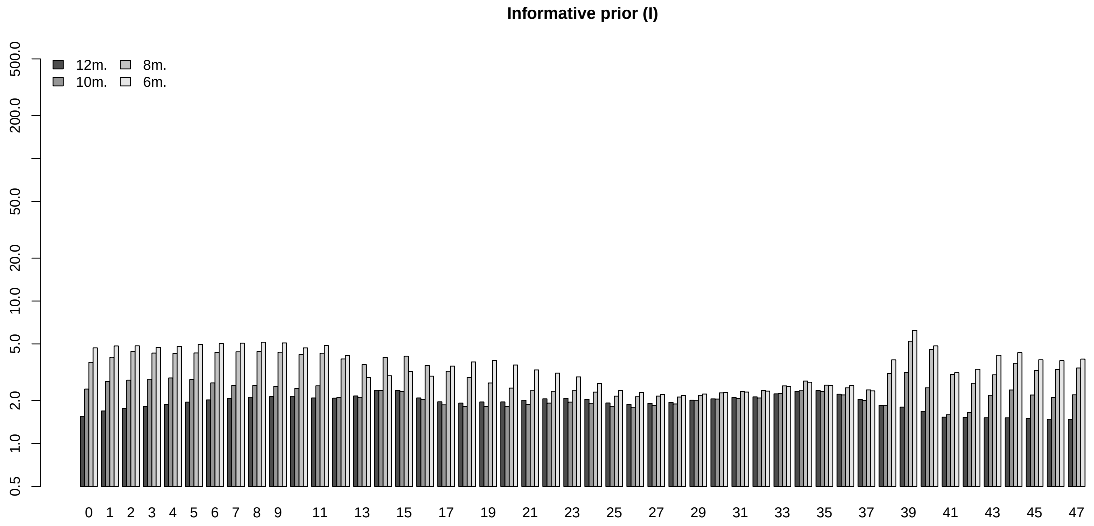


Figure 3.12: Using the informative prior: quality of the predictions (MAPE in %) averaged for each instant (all the 180 or so days within the prediction time-window are used for those averages), with an estimation period ranging from 12 to 6 months. The ordinate axis is in log-scale. Each shade of gray corresponds to a different scenario.

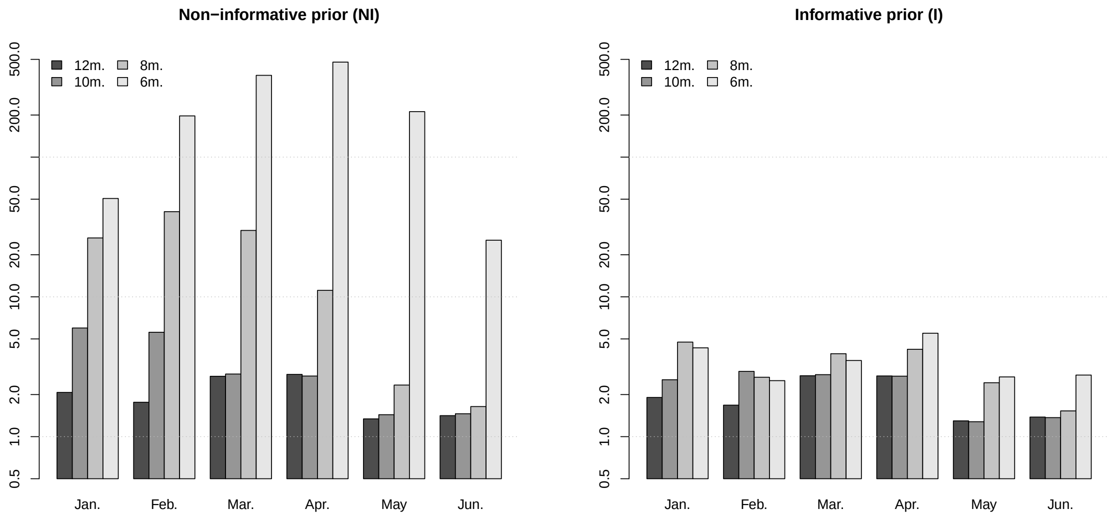


Figure 3.13: Quality of the predictions (MAPE in %) averaged for each month (all the instants of the 30 or so days within each month are used to compute these averages), with an estimation period ranging from 12 to 6 months using the non informative prior (left) and using the informative prior (right). The ordinate axis is in log-scale. Each shade of gray corresponds to a different scenario.

3.6.1 Technical Lemmas

Definition 3.5 (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 3.6 (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

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

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 3.7 (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 3.8. *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. ■

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

The different steps of the MCMC algorithm we used to (approximately) simulate $(\theta_1, \ldots, \theta_M)$ according to the posterior distribution $\pi(\theta|y)$ that corresponds to the non informative prior we presented earlier are given in Algorithm 3.1. 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 given hereafter. Lemma 3.8 is a key element to these justifications.

Choosing Σ_{MH}. The covariance matrix Σ_{MH} used in the 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).

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 3.8, it is then straightforward to see that the full conditional distribution of α is Gaussian

$$\alpha|\theta\backslash\alpha, y \sim \mathcal{N}(\mu^\alpha, \Sigma^\alpha) \tag{3.14}$$

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

$$\beta|\theta\backslash\beta, y \sim \mathcal{N}(\mu^\beta, \Sigma^\beta, B_+^{d_\beta}(0,1)) \tag{3.15}$$

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

$$\gamma|\theta\backslash\gamma, y \sim \mathcal{N}(\mu^\gamma, \Sigma^\gamma) \tag{3.16}$$

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 (3.7).

Algorithm 3.1 (MCMC algorithm for the estimation of the posterior distribution, using the non informative prior.)**.**

For $t = 1$

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

For $t \geqslant 1$

1. Simulate σ^2_{t+1} cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, y)$ i.e.
$$\sigma^2_{t+1} \sim \mathcal{IG}\left(\frac{N}{2}, \frac{1}{2}\|y - f(\eta)\|_2^2\right).$$
2. Simulate γ_{t+1} cond. to $(\alpha_t, \beta_t, u_t, \sigma^2_{t+1}, y)$ i.e.
$$\gamma_{t+1} \sim \mathcal{N}\left(\mu^{\gamma}_{t+1}, \Sigma^{\gamma}_{t+1}\right).$$
3. Simulate b_{t+1} cond. to $(\alpha_t, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y)$ i.e.
$$\beta_{t+1} \sim \mathcal{N}\left(\mu^{\beta}_{t+1}, \Sigma^{\beta}_{t+1}, B_+^{d_\beta}(0,1)\right).$$
4. Simulate a_{t+1} cond. to $(\beta_{t+1}, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y)$ i.e.
$$\alpha_{t+1} \sim \mathcal{N}\left(\mu^{\alpha}_{t+1}, \Sigma^{\alpha}_{t+1}\right).$$
5. Simulate $\delta_t \sim \mathcal{N}(0, \Sigma_{\mathrm{MH}})$ and set $\widetilde{u}_t \leftarrow u_t + \delta_t$.
6. Compute
$$\rho(u_t, \widetilde{u}_t) = \frac{\pi(\widetilde{u}_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, y)}{\pi(u_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, y)}.$$
7. Simulate $v_t \sim \mathcal{U}[0, 1]$. If $v_t < \rho(u_t, \widetilde{u}_t)$ then set $u_{t+1} \leftarrow \widetilde{u}_t$ else set $u_{t+1} \leftarrow u_t$.

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

The different steps of the MCMC algorithm we used to (approximately) simulate $(\theta_1, \dots, \theta_M)$ according to the posterior distribution $\pi(\theta|y)$ that corresponds to the non informative prior we presented earlier are given in Algorithm 3.2. 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 are given hereafter. To derive these full conditional distributions, we will make use of the technical Lemmas 3.6, 3.7 and 3.8 presented earlier.

Choosing Σ_{MH}. As for Algorithm 3.1, the covariance matrix Σ_{MH} used in the 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)

Algorithm 3.2 (MCMC algorithm for the estimation of the posterior distribution, using the informative prior.)**.**

```
For $t = 1$
        Initialise $\theta_1$ such that $\pi(\theta_1|y) \neq 0$.
For $t \geqslant 1$
    1. Simulate $\sigma^2_{t+1}$ cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, k_t, l_t, q_t, r_t, y)$ i.e.
           $$\sigma^2_{t+1} \sim \mathcal{IG}\left(\frac{N}{2}, \frac{1}{2}\|y - f(\eta)\|_2^2\right)$$
    2. Simulate $r_{t+1}$ cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, l_t, q_t, y)$ i.e.
           $$r_{t+1} \sim \mathcal{G}\left(a_r + \frac{d}{2}, b_r + \frac{1}{2}\sum_{i=1}^{d}(k_i - q)^2\right)$$
    3. Simulate $q_{t+1}$ cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, l_t, r_{t+1}, y)$ i.e.
           $$q_{t+1} \sim \mathcal{N}\left([\sigma_q^{-2} + rd]^{-1}(\sigma_q^{-2} + r\sum_{i=1}^{d} k_i), [\sigma_q^{-2} + rd]^{-1}\right)$$
    4. Simulate $l_{t+1}$ cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, q_{t+1}, r_{t+1}, y)$ i.e.
           $$l_{t+1} \sim \mathcal{G}\left(a_l + \frac{d}{2}, b_l + \frac{1}{2}(\eta_t - K\mu_t^{\mathcal{A}})'(\Sigma^{\mathcal{A}})^{-1}(\eta_t - K\mu_t^{\mathcal{A}})\right)$$
    5. Simulate $k_{t+1}$ cond. to $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y)$ i.e.
           $$k_{t+1} \sim \mathcal{N}\left(\mu^k_{t+1}, \Sigma^k_{t+1}\right)$$
    6. Simulate $\gamma_{t+1}$ cond. to $(\alpha_t, \beta_t, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y)$ i.e.
           $$\gamma_{t+1} \sim \mathcal{N}\left(\mu^g_{t+1}, \Sigma^g_{t+1}\right)$$
    7. Simulate $\beta_{t+1}$ cond. to $(\alpha_t, \gamma_{t+1}, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y)$ i.e.
           $$\beta_{t+1} \sim \mathcal{N}\left(\mu^b_{t+1}, \Sigma^b_{t+1}, B_+^{d_\beta}(0,1)\right)$$
    8. Simulate $\alpha_{t+1}$ cond. to $(\beta_{t+1}, \gamma_{t+1}, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y)$ i.e.
           $$\alpha_{t+1} \sim \mathcal{N}\left(\mu^a_{t+1}, \Sigma^a_{t+1}\right)$$
    9. Simulate $\delta_t \sim \mathcal{N}(0, \Sigma_{\text{MH}})$ and set $\widetilde{u}_t \leftarrow u_t + \delta_t$
   10. Compute
           $$\rho(u_t, \widetilde{u}_t) = \frac{\pi(\widetilde{u}_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y)}{\pi(u_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y)}.$$
   11. Simulate $v_t \sim \mathcal{U}[0, 1]$. If $v_t < \rho(u_t, \widetilde{u}_t)$ then set $u_{t+1} \leftarrow \widetilde{u}_t$ else set $u_{t+1} \leftarrow u_t$.
```

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 3.6 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 3.6 (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 (3.14).

3. With the help of Lemma 3.6 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 (3.15), and combining the prior and the likelihood contributions together with Lemma 3.6 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 (3.16), and combining the prior and the likelihood contributions together with Lemma 3.6 we obtain the Gaussian distribution

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

where

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

Full conditional distribution of k. 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 3.6, 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 (3.6).

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.

4 An application of particle filters to electricity load forecasting

In this paper, we are interested in the online prediction of the electricity load, within the Bayesian framework of dynamic models. We offer a review of sequential Monte Carlo methods, and provide the calculations needed for the derivation of so-called particles filters. We also discuss the practical issues arising from their use, and some of the variants proposed in the literature to deal with them, giving detailed algorithms whenever possible for an easy implementation. We propose an additional step to help make basic particle filters more robust with regard to outlying observations. Finally we use such a particle filter to estimate a state-space model that includes exogenous variables in order to forecast the electricity load for the customers of the French electricity company Électricité de France and discuss the various results obtained.

4.1 Introduction

Let $\{X_n\}_{n\geqslant 0}$ and $\{Y_n\}_{n\geqslant 0}$ be $\mathcal{X} \subset \mathbb{R}^{n_x}$ and $\mathcal{Y} \subset \mathbb{R}^{n_y}$-valued stochastic processes defined on a measurable space. The observations $\{Y_n\}_{n\geqslant 0}$ are assumed conditionally independent given the hidden Markov process $\{X_n\}_{n\geqslant 0}$ most often referred to as the states of the model, and are characterised by the conditional density $g_n^\theta(y_n|x_n)$. We denote the initial density of the state as $\mu^\theta(x_0)$ and the Markov transition density from time $n-1$ to time n as $f_n^\theta(x_n|x_{n-1})$. The superscript θ on these densities is the parameter of the model, that belongs to an open set $\Theta \subset \mathbb{R}^{n_\theta}$. The model can be summarised (using practical and common if not exactly rigorous notations) as

$$X_0 \sim \mu^\theta(\cdot), \quad X_n|(X_{n-1} = x_{n-1}) \sim f_n^\theta(\cdot|x_{n-1}) \tag{4.1}$$

$$Y_n|(X_n = x_n) \sim g_n^\theta(\cdot|x_n). \tag{4.2}$$

Within the Bayesian framework, equations (4.1) specify the prior on the states of the model whose likelihood is defined via (4.2).

Notice here that we restrict ourselves to models with independent observations, but that the framework can easily be extended to include dependent observations if need be. The class of dynamic models we consider, known as general state-space models or hidden Markov models (HMM) in the literature and whose typical representation is given in Figure 4.1, includes many non linear and non

Gaussian time series models such as

$$X_{n+1} = F_n(X_n, V_{n+1}) \tag{4.3}$$

$$Y_n = G_n(X_n, W_n) \tag{4.4}$$

where $\{V_n\}_{n \geqslant 1}$ and $\{W_n\}_{n \geqslant 0}$ are independent sequences of independent random variables and $\{F_n\}_{n \geqslant 1}$ and $\{G_n\}_{n \geqslant 1}$ are sequences of (possibly non linear) functions. Such models find applications in many fields including time-series forecasting (Dordonnat, 2009), biostatistics (Rossi, 2004; Vavoulis et al., 2012), econometrics (Chopin et al., 2012; Johansen et al., 2008; Liu and West, 2001), telecommunications (Lee et al., 2010), object tracking (Gilks and Berzuini, 2001; Gustafsson et al., 2002; Karlsson, 2005; Rui and Chen, 2001), etc.

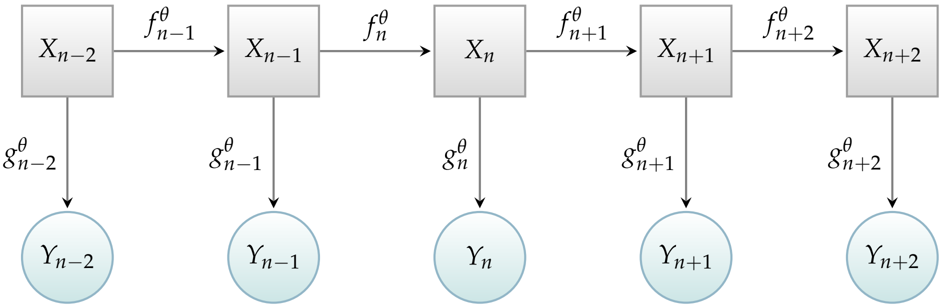


Figure 4.1: A generic hidden Markov Model (HMM).

When the parameter θ is known, on-line inference about the state process given the observations is a so-called optimal filtering problem. For simple models such as the linear Gaussian state-space model the problem can be solved exactly using the standard Kalman filter (see for example Durbin and Koopman, 2001), and the case of a finite state-space also allows for explicit calculations. For non linear models, the Extended Kalman filter is often used and relies on the approximation of the first derivative of F_n, although good performances are not guaranteed theoretically. Another technique is the so-called Unscented Kalman filter (see Wan and van der Merwe, 2000, for the comprehensive details) which makes use of the unscented transformation to deal with the non linearity of the system.

For our application, we are interested in the on-line prediction of the french electricity load through the estimation (and prediction) of a dynamic model and choose to consider Sequential Monte Carlo (SMC) methods also known as particle methods instead. SMC methods are a class of sequential simulation-based algorithms which aim at approximating the posterior distributions of interest. They represent a popular alternative to Kalman filters (Kantas et al., 2009) since they are often easy to implement, apply to non linear non Gaussian models, and have been demonstrated to yield accurate estimates (Doucet et al., 2001; Liu, 2008).

In Section 4.2 we introduce the key concepts behind sequential Monte Carlo methods. In Section 4.3 we first derive the algorithm for a basic particle filter and discuss common practical issues. We then review the main techniques appearing in the literature to deal with these issues and we also propose a new additional step to help make particle filters more robust with regard to outlying observations. Finally, we propose a new nonlinear dynamic model for the electricity load in Section 4.4 and use a particle filter to estimate this model. We compare the predictions we obtain to operational predictions and show that our model remains competitive, even though its definition is simpler than that of the model studied in Dordonnat et al. (2008).

4.2 TOWARDS SEQUENTIAL MONTE CARLO

Let us first assume that the parameter θ is known: the model with θ unknown will be discussed later in Section 4.3.7.

4.2.1 Posterior distribution

Given equations (4.1) and (4.2), the posterior distribution of the states given the observations is

$$\pi^{\theta}(x_{0:n}|y_{0:n}) \propto \underbrace{\prod_{k=1}^{n} g_k^{\theta}(y_k|x_k)}_{n \text{ likelihoods}} \cdot \underbrace{\prod_{k=1}^{n} f_k^{\theta}(x_k|x_{k-1})}_{n \text{ transition densities}} \cdot \underbrace{\mu^{\theta}(x_0)}_{\text{initial density}} \quad . \tag{4.5}$$

From equation (4.5), three distinct goals might be pursued (see for example Cappé et al., 2010; Chen, 2003)

Filtering: the aim of filtering is to estimate the distribution of the state X_n conditionally to the observations up to time n, i.e. $y_{0:n}$.

Smoothing: the aim of smoothing is to estimate the distribution of the state X_n conditionally to the observations up to time n' (with $n' \geqslant n$), i.e. $y_{0:n'}$. Note that $\pi^{\theta}(x_n|y_{0:n})$ is both a filtered and a smoothed distribution.

Predicting: the aim of predicting is to estimate the distribution of the state $X_{n+\tau}$ (with an horizon $\tau > 0$) conditionally to the observations up to time n, i.e. $y_{0:n}$. From there, using (4.2), it is easy to forecast the upcoming observation $Y_{n+\tau}$ which is usually the real target. When not explicitly mentioned, the horizon considered for prediction will be $\tau = 1$.

To summarise, given the available observations, filtering focuses on the current state, smoothing focuses on the past states, and predicting focuses on the future states. Our goal being the online prediction of the electricity load, we chose to focus on predicting and filtering, since the filtered distribution of the state at time n is needed to produce forecasts for time $n + \tau$: ultimately, smoothing only refines the estimation of past states over time, without influencing the quality of the online prediction, and is therefore not needed to achieve our goal.

4.2.2 The Monte Carlo toolbox for Bayesian inference

Markov Chains Monte Carlo

MCMC methods (see for example Marin and Robert, 2007; Robert, 1996; Robert and Casella, 2004) certainly represent a viable estimation procedure: most of the time, nothing really prevents the exploration via MCMC of the posterior distribution derived in (4.5) from the prior and the likelihood given in (4.1) and (4.2). From a practical point of view however, MCMC methods are most likely not the optimal tool: the addition of a new observation y_{n+1} from the model forces the overall re-estimation of the smoothed distribution of the states $\pi^\theta(x_{0:n+1}|y_{0:n+1})$ even when we are interested only in the last marginal of this distribution i.e. the filtered distribution $\pi^\theta(x_{n+1}|y_{0:n+1})$. The MCMC estimation is thus not recursive (with regard to the time index) in the sense that the filtered distribution $\pi^\theta(x_{n+1}|y_{0:n+1})$ at time $n+1$ cannot be computed from the previous filtered distribution $\pi^\theta(x_n|y_{0:n})$ at time n using MCMC methods, which is a major drawback given the computationally expensive nature of these methods.

Notice also that even though designing the MCMC algorithm can be simple in some cases, the dimension of the space explored grows linearly with the time index making the assessment of the convergence of the produced Markov chains all the more complicated.

Sequential Monte Carlo

SMC methods provide a viable and popular alternative to MCMC methods for the Bayesian online estimation of dynamic models. Particle methods are recursive by nature (thus computationally cheaper than MCMC) and similar in some ways to the Kalman filter approach. Particle methods essentially draw their strength from the immediate calculations that we show below

$$\begin{aligned}
\pi^\theta(x_{0:n}|y_{0:n}) &= \frac{\pi^\theta(y_{0:n}|x_{0:n})\pi^\theta(x_{0:n})}{\pi^\theta(y_{0:n})} = \frac{\pi^\theta(y_n,y_{0:n-1}|x_{0:n})\pi^\theta(x_{0:n})}{\pi^\theta(y_n,y_{0:n-1})}\\
&= \frac{\pi^\theta(y_n|y_{0:n-1},x_{0:n})\pi^\theta(y_{0:n-1}|x_{0:n})\pi^\theta(x_{0:n})}{\pi^\theta(y_n|y_{0:n-1})\pi^\theta(y_{0:n-1})}\\
&= \frac{\pi^\theta(y_n|y_{0:n-1},x_{0:n})\pi^\theta(x_{0:n}|y_{0:n-1})\pi^\theta(y_{0:n-1})\pi^\theta(x_{0:n})}{\pi^\theta(y_n|y_{0:n-1})\pi^\theta(y_{0:n-1})\pi^\theta(x_{0:n})}\\
&= \frac{\pi^\theta(y_n|x_{0:n})\pi^\theta(x_n|x_{0:n-1},y_{0:n-1})}{\pi^\theta(y_n|y_{0:n-1})}\cdot\pi^\theta(x_{0:n-1}|y_{0:n-1})
\end{aligned}$$

i.e. with the notations we introduced earlier:

$$\begin{aligned}
\pi^\theta(x_{0:n}|y_{0:n}) &= \frac{g_n^\theta(y_n|x_n)f_n^\theta(x_n|x_{n-1})}{\pi^\theta(y_n|y_{0:n-1})}\cdot\pi^\theta(x_{0:n-1}|y_{0:n-1}) \qquad (4.6)\\
&\propto g_n^\theta(y_n|x_n)f_n^\theta(x_n|x_{n-1})\cdot\pi^\theta(x_{0:n-1}|y_{0:n-1}).
\end{aligned}$$

The recursive equation (4.6) plays a central role in the definition of all particle methods. An integrated version of this equation is most often presented to emphasise the direct connection between two consecutive filtered distributions:

$$\begin{aligned}
\pi^\theta(x_n|y_{0:n}) &= \int \pi^\theta(x_{0:n}|y_{0:n})\,\mathrm{d}x_{0:n-1} \qquad (4.7)\\
&\propto g_n^\theta(y_n|x_n)\int f_n^\theta(x_n|x_{n-1})\cdot\pi^\theta(x_{n-1}|y_{0:n-1})\,\mathrm{d}x_{n-1}.
\end{aligned}$$

The main idea behind particle filters is to make extensive use of equation (4.6) to compute sequential Monte Carlo approximations of the posterior distributions of interest, in our case, the sequence of filtered distributions. The general procedure is simple enough and mimics the iterative prediction-correction structure of any Kalman filter. A each time n the filtered density $\pi^\theta(x_n|y_{0:n})$ can be approximated by the empirical distribution of a large sample of M ($M >> 1$) weighted random samples termed particles. The weighted particles evolve over time: they follow the prior dynamic distribution of the model and get re-adjusted as soon as observations become available. At time n, the two basic steps (a lot of refinements are possible that we will discuss later on) of particle filters are the following:

Prediction: given particles distributed along density $\pi^\theta(x_{n-1}|y_{0:n-1})$, we simulate new particles distributed along density $\pi^\theta(x_n|y_{0:n-1})$ with the help of the transition density $f_n^\theta(x_n|x_{n-1})$.

Correction: we re-weight these particles distributed along density $\pi^\theta(x_n|y_{0:n-1})$ depending on the observation y_n with the help of (4.6) to approximate the distribution $\pi(x_n|y_{0:n})$.

Particle filters essentially combine two basic techniques: Monte Carlo integration and importance sampling. We offer a brief overview of these two numerical integration techniques to show how closely they are connected to SMC.

4.2.3 Numerical approximation

Monte Carlo integration allows the estimation of integrals of the form

$$I = \mathbb{E}^\pi[h(X)] = \int h(x)\pi(x)\,\mathrm{d}x, \tag{4.8}$$

where π is a probability density and where $h \in \mathrm{L}^1(\pi)$. This method is often used to numerically approximate the expectation of a random variable whose density is π or a moment of higher order.

Importance sampling

The importance sampling method provides a way to estimate (4.8), that extends the results given above. We now assume that a probability density q (the so-called importance density) is available from which we can simulate, and such that the support of π is included in that of q. Re-writing

$$I = \int h(x)\pi(x)\,\mathrm{d}x = \int \frac{h(x)\pi(x)}{q(x)} q(x)\,\mathrm{d}x,$$

and remembering that $h \in \mathrm{L}^1(\pi)$, it is easy to see that $h\pi/q \in \mathrm{L}^1(q)$. Hence, it becomes clear that it is in fact possible to use Monte Carlo integration with q instead of π, and this for any choice of q satisfying to the support condition.

Given $X^1, \dots, X^M$ i.i.d. random variables with probability density q, the importance sampling estimator of I is defined by

$$\widehat{I}_M^{\mathrm{IS}}(q) = \frac{1}{M}\sum_{j=1}^{M} \frac{\pi(X^j)h(X^j)}{q(X^j)} = \frac{1}{M}\sum_{j=1}^{M} \widetilde{w}^j h(X^j) \tag{4.9}$$

where

$$\widetilde{w}^j = \frac{\pi(X^j)}{q(X^j)}. \tag{4.10}$$

For $h \in \mathrm{L}^1(\pi)$, it is then straightforward that:

$$\mathbb{E}^q[\widehat{I}_M^{\mathrm{IS}}(q)] = I, \qquad \widehat{I}_M^{\mathrm{IS}}(q) \xrightarrow{a.s} I.$$

A drawback of the importance sampling presented above resides in the fact that even though it allows for an estimation of integrals such as (4.8), it does not provide an approximation of the density π by the empirical distribution of the particles X^j weighted by $\widetilde{w}^j$ because the mean of the weights $\frac{1}{M}\sum_{j=1}^{M}\widetilde{w}^j$ is not equal to 1. An approximation similar to that obtained for the Monte Carlo estimator is nevertheless possible given a slight modification to the original importance sampling technique.

Self-normalised importance sampling

Using the same notations as in the previous section, since π is a probability density, we can also write

$$I = \int h(x)\pi(x)\,\mathrm{d}x = \frac{\displaystyle\int \frac{h(x)\pi(x)}{q(x)} q(x)\,\mathrm{d}x}{\displaystyle\int \frac{\pi(x)}{q(x)} q(x)\,\mathrm{d}x}.$$

A self-normalised importance sampling version of the previous estimator can thus be formulated. Given $X^1, \ldots, X^M$ i.i.d. random variables with probability density q, the self-normalised importance sampling estimator of I is defined by

$$\widehat{I}_M^{\mathrm{SNIS}}(q) = \frac{\displaystyle\sum_{j=1}^{M} \frac{h(X^j)\pi(X^j)}{q(X^j)}}{\displaystyle\sum_{j=1}^{M} \frac{\pi(X^j)}{q(X^j)}} = \sum_{j=1}^{M} w^j h(X^j), \tag{4.11}$$

where, remembering the definition of the weights $\widetilde{w}^j$ given in (4.10), we define the self-normalised weights as

$$w^j = \frac{\widetilde{w}^j}{\sum_{k=1}^{M} \widetilde{w}^k}. \tag{4.12}$$

For $h \in \mathrm{L}^1(\pi)$, even though the estimate is shown to be biased, it achieves consistency (see for example Geweke, 1989)

$$\mathbb{E}^q[\widehat{I}_M^{\mathrm{SNIS}}(q)] \to I, \qquad \widehat{I}_M^{\mathrm{SNIS}}(q) \xrightarrow{a.s} I.$$

For q such that $\int \frac{\pi^2(x)}{q(x)}\,\mathrm{d}x < +\infty$, and $h \in \mathrm{L}^2(\pi^2/q)$, i.e. such that $\int \frac{h^2(x)\pi^2(x)}{q(x)}\,\mathrm{d}x < +\infty$, let us denote

$$\sigma_q^2 = \mathrm{Var}^q[h(X^1)\pi(X^1)/q(X^1)],$$

the variance of $h(X)\pi(X)/q(X)$ when the distribution of X has density q.

We then have (see Geweke, 1989, for the proof)

$$\sqrt{n}(\widehat{I}_M^{\mathrm{SNIS}} - I) \xrightarrow{d} N(0, \sigma_q^2).$$

Observe now that using the self-normalised weights w^j as defined in (4.12), (4.11) can be reformulated as:

$$\widehat{I}_M^{\mathrm{SNIS}}(q) = \int h(x)\mu_M^{\mathrm{SNIS}}(\mathrm{d}x), \tag{4.13}$$

where

$$\mu_M^{\mathrm{SNIS}}(\cdot) = \sum_{j=1}^{M} w^j \delta(X^j, \cdot) \tag{4.14}$$

is a probability measure that yet again approximates the measure of density π consistently.

Application to dynamic models

We describe the application of self-normalised importance sampling to estimate a sequence of integrals that involve the posterior distribution (4.5) and that are of the form

$$\begin{aligned} I_n &= \int h(x_n)\pi^\theta(x_{0:n}|y_{0:n})\,\mathrm{d}x_{0:n} \\ &= \int h(x_n)\pi^\theta(x_n|y_{0:n})\,\mathrm{d}x_n. \end{aligned}$$

We use the self-normalised importance sampling estimator defined in (4.11), with $\pi(x) = \pi^\theta(x_{0:n}|y_{0:n})$ and $q(x) = q(x_{0:n}|y_{0:n})$. Given M particles $X^1_{0:n}, \ldots, X^M_{0:n}$, i.i.d. with probability density $q^\theta(x_{0:n}|y_{0:n})$, we will approximate I_n by

$$\widehat{I}^{\mathrm{PF}}_{n,M} = \sum_{j=1}^{M} w^j_n h(X^j_n),$$

where mimicking the definitions given (4.10) and (4.12) we define

$$w^j_n = \frac{\widetilde{w}^j_n}{\sum_{k=1}^{M}\widetilde{w}^k_n}, \tag{4.15}$$

with

$$\widetilde{w}^j_n = \frac{\pi^\theta(X^j_{0:n}|y_{0:n})}{q^\theta(X^j_{0:n}|y_{0:n})}. \tag{4.16}$$

Note that to alleviate the notational burden, we voluntarily omit the dependence of the importance weights on the parameter θ, and will do so for the remainder of the chapter when no confusion is possible.

A convenient form of importance density

Let us consider an importance density q that can be factorised as follows:

$$\begin{aligned} q^\theta(x_{0:n}|y_{0:n}) &= q^\theta(x_n|y_{0:n-1}, x_{0:n})q^\theta(x_{0:n-1}|y_{0:n-1}) \\ &= q^\theta(x_0|y_0)\prod_{k=1}^{n} q^\theta(x_k|y_{0:k-1}, y_{0:k}). \end{aligned} \tag{4.17}$$

It is now easy to see, using (4.6), that the weights $\widetilde{w}^\theta_n(X^j_{0:n})$ can be updated recursively via

$$\begin{aligned} \widetilde{w}^j_n = \frac{\pi^\theta(X^j_{0:n}|y_{0:n})}{q^\theta(X^j_{0:n}|y_{0:n})} &= \frac{g^\theta_n(y_n|X^j_n)f^\theta_n(X^j_n|X^j_{n-1})\pi^\theta(X^j_{0:n-1}|y_{0:n-1})}{\pi^\theta(y_n|y_{0:n-1})q^\theta(X^j_n|X^j_{0:n-1}, y_{0:n})q^\theta(X^j_{0:n-1}|y_{0:n-1})} \\ &= \widetilde{w}^j_{n-1}\frac{g^\theta_n(y_n|X^j_n)f^\theta_n(X^j_n|X^j_{n-1})}{\pi^\theta(y_n|y_{0:n-1})q^\theta(X^j_n|X^j_{0:n-1}, y_{0:n})}. \end{aligned} \tag{4.18}$$

where $\pi^\theta(y_n|y_{0:n-1})$ does not depend on the index j, and need not be computed at all since the weights w^j_n featured in the estimator are the self-normalised version of the weights $\widetilde{w}^j_n$ (the constant vanishes after the self-normalisation). Note that w^j_{n-1} can be substituted to $\widetilde{w}^j_{n-1}$ in the recursive update (4.18) for the very same reason.

Equation (4.18) lies at the very core of all the particle filters in general, some variants of which we describe in the next section. It summarises, by itself, the edge that SMC methods have over MCMC methods in general in the context of dynamic models: it allows for sequential recursive estimations and predictions. At each time step, two things only are required to estimate the quantity of interest: simulations from the importance density q^θ (the choice of which shall be discussed) and the update of the particles' weights via the computation of (4.18).

4.3 Particle filters

From this point on, we adopt the convention that whenever the index j is used, we mean "for all $j = 1, \dots, M$". We present SMC methods designed to approximate the sequence of filtered distributions $\pi^\theta(x_n|y_{0:n})$: at the end of each time step n, the particle filters discussed hereafter return M particles X_n^j with weights w_n^j that can be used to approximate for instance

- the filtered distribution $\pi^\theta(x_n|y_{0:n})$ by the finite mixture of weighted Dirac masses

$$\widehat{\pi}(\mathrm{d}x_n|y_{0:n}) = \sum_{j=1}^M w_n^j \delta(X_n^j, \mathrm{d}x_n),$$

- integrals such as $I_n = \displaystyle\int h(x_n)\pi(x_n|y_{0:n})\,\mathrm{d}x_n$, with $h \in L^1(\pi(\cdot|y_{0:n}))$, by

$$\widehat{I}_{n,M} = \sum_{j=1}^M w_n^j h(X_n^j).$$

4.3.1 Sequential Importance Sampling (SIS)

Conception

The SIS filter (sometimes also called Bayesian Importance Sampling) is a direct application of the calculations shown in the previous section: it relies solely upon the sequential use of the self-normalised importance sampling technique. The details are given in Algorithm 4.1.

Algorithm 4.1 (Sequential Importance Sampling (SIS) for filtering)**.**

```
At time $n = 0$
    1. Sample $X_0^j \sim q^\theta(x_0|y_0)$.
    2. Compute $\widetilde{w}_0^j = \dfrac{g_0^\theta(y_0|X_0^j)\mu^\theta(X_0^j)}{q^\theta(X_0^j|y_0)}$ and set $w_0^j \leftarrow \dfrac{\widetilde{w}_0^j}{\sum_{k=1}^M \widetilde{w}_0^k}$.
At time $n \geqslant 1$
    1. Sample $X_n^j \sim q^\theta(x_n|x_{0:n-1}, y_{0:n})$.
    2. Compute $\widetilde{w}_n^j = w_{n-1}^j \dfrac{g_n^\theta(y_n|X_n^j)f_n^\theta(X_n^j|X_{n-1}^j)}{q^\theta(X_n^j|X_{0:n-1}^j, y_{0:n})}$ and set $w_n^j \leftarrow \dfrac{\widetilde{w}_n^j}{\sum_{k=1}^M \widetilde{w}_n^k}$.
```

At each time step, new particles are first simulated conditionally to the old ones to represent the predictive distribution of the upcoming state and, as the observation becomes available, their weights then get readjusted to represent the filtered distribution.

Prediction

The estimation of the predicted distribution $\pi^\theta(x_{n+\tau}|y_{0:n})$ ($\tau \geqslant 1$) can also be computed from the estimation of the filtered distribution up to time n. The principle, described for instance in Doucet (1998), is identical in essence to that developed in Durbin and Koopman (2001) for Kalman filters. Since the observations at times $n+1, \dots, n+\tau$ are not yet available, no correction may take place after the predictions of the state that involve the transition densities $f^\theta_{n+\tau}, \dots, f^\theta_{n+1}$: formally, the terms $g^\theta_{n+\tau}, \dots, g^\theta_{n+1}$ vanish. The details are given in Algorithm 4.2. Observe that in this case, the importance density $q^\theta(x_{n+\tau}|x_{0:n+\tau-1}, y_{0:n})$ needs to be chosen so as not to involve the yet unknown values of the upcoming observations $y_{n+1:n+\tau}$.

Algorithm 4.2 (Sequential Importance Sampling (SIS) for predicting)**.**

At time $n \geqslant 0$**, for** $\tau = 1, \dots$

1. Sample $X^j_{n+\tau} \sim q^\theta(x_{n+\tau}|x_{0:n+\tau-1}, y_{0:n})$.
2. Compute $\widetilde{w}^j_{n+\tau} = w^j_{n+\tau-1} \dfrac{f^\theta_{n+\tau}(X^j_{n+\tau}|X^j_{n+\tau-1})}{q^\theta(X^j_{n+\tau}|X^j_{0:n+\tau-1}, y_{0:n})}$ and set $w^j_{n+\tau} \leftarrow \dfrac{\widetilde{w}^j_{n+\tau}}{\sum_{k=1}^M \widetilde{w}^k_{n+\tau}}$.

Missing observations

When dealing with a missing observation, the SIS filter requires little modification: when observation Y_n is missing, the corresponding state X_n is predicted using Algorithm 4.2 since $\pi^\theta(x_n|y_{0:n-1})$ is the only accessible density under such circumstances. This leads to Algorithm 4.3.

Algorithm 4.3 (Sequential Importance Sampling (SIS) for filtering with missing observations)**.**

At time $n \geqslant 0$**, if observation** Y_n **is missing**

1. Sample $X^j_n \sim q^\theta(x_n|x_{0:n-1}, y_{0:n-1})$.
2. Compute $\widetilde{w}^j_n = w^j_{n-1} \dfrac{f^\theta_n(X^j_n|X^j_{n-1})}{q^\theta(X^j_n|X^j_{0:n-1}, y_{0:n-1})}$ and set $w^j_n \leftarrow \dfrac{\widetilde{w}^j_n}{\sum_{k=1}^M \widetilde{w}^k_n}$.

Comments

The major drawback of the SIS filter comes from the fact that the distribution of the weights degenerates, with the variance of the importance weights increasing over time (see Doucet et al., 2000) meaning that the estimated distributions become less and less unreliable: after a few iterations, all but one of the normalised importance weights are close to zero. An important fraction of the calculations involved in the algorithm is thus dedicated to particles whose contributions to the estimation are almost null, making the SIS particle filter an impractical estimation procedure at best.

4.3.2 Monitoring the degeneracy

To alleviate the degeneracy problem that we outlined, additional steps are traditionally implemented into Algorithm 4.1. Since adding these new steps comes at a non negligible computational cost, it is important to somehow monitor how badly the weight distribution degenerates at a given time step, because it is usually interesting to ignore the degeneracy problem unless it reaches a given threshold.

A popular rule of thumb, first introduced in Kong et al. (1994) and later copiously reprised in the literature (see for instance Chen, 2003; Doucet et al., 2000; Liu, 2008), is to consider the so-called effective sample size based on the normalised weights w_n^j at time step n and defined by

$$\frac{M}{1 + \mathrm{Var}^{q^\theta(\cdot|y_{0:n})}[w_n^1]}.$$

This quantity is usually numerically approximated by the following estimate

$$\mathrm{ESS}(n) = \frac{1}{\sum_{k=1}^M (w_n^k)^2}. \tag{4.19}$$

It ranges from M (reached when all the particles share equal weights of value 1) to $1/M$ (reached when a single particle is given the whole probability mass of the sample, with a weight of 1).

A related degeneracy measure is the coefficient of variation (found in Kong et al., 1994; Liu and Chen, 1995), ranging from 0 to $\sqrt{M-1}$, that is given by

$$\mathrm{CV}(n) = \sqrt{\frac{1}{M}\sum_{k=1}^{M}(Mw_n^k - 1)^2}, \tag{4.20}$$

and satisfies to

$$\mathrm{ESS}(n) = \frac{M}{1 + \mathrm{CV}(n)^2}. \tag{4.21}$$

The Shannon entropy of the importance weights, ranging from $\log M$ to 0, is sometimes also mentionned. It is defined by

$$\mathcal{E}(n) = -\sum_{k=1}^{M} w_n^k \log w_n^k. \tag{4.22}$$

Cornebise (2009) recently proved that the criteria (4.20) and (4.22) are estimators of the χ^2-divergence and the Kullback-Leibler divergence between two distributions which are associated with the importance and target densities of the particle filter.

The evaluation of one (or more) of these criteria is introduced at each time step, with the additional procedures that we discuss next taking place if and only if the criterion reaches a certain fixed threshold so as to reduce the additional computational burden. The most common threshold found in the literature is $\mathrm{ESS}(n) < 0.5M$. Examples illustrating the behaviours of these criteria are given later in Figures 4.3, 4.4 and 4.5.

4.3.3 Resample step

A resampling step is most often introduced into Algorithm 4.2 to help and fight the degeneracy problem. The aim of this resampling step is to favour the living of the interesting particles (the ones with more important weights, that are more representative of the targeted distribution) and encourage the dying of the not so interesting particles so as to focus the computational effort upon particles that matter most for the estimation. The resampling method has to be carefully chosen, in particular it should not introduce any bias in the final estimate as mentioned in Doucet et al. (2000)

During this new step, particles are resampled according to their weights: a particle with an important weight is more likely to appear (and ”survive”) in the new sample generated, possibly more than once, whereas a particle the weight of which is close to zero is more likely not to be drawn at all (and “die“) from a given time step to the next.

Chen (2003) mentions that there are a few resampling schemes available in the literature. It is important to note that even though resampling might alleviate the degeneracy problem, it also brings extra random variation to the samples of particles. As a consequence, the filtered quantities of interest should preferably be computed before resampling and not after. We only present the details of the multinomial and residual resampling schemes.

Multinomial resampling

Multinomial resampling is the most popular resampling scheme, most likely because it is the easiest to both understand and implement: at a given time step, it suffices to simulate a discrete random variable which takes values X_n^k with probability w_n^k. The details of multinomial resampling are given in Algorithm 4.4 where only the new step is described.

Algorithm 4.4 (Multinomial resampling step).

```
At time $n \geqslant 0$
    3. Sample $Z_n^j \sim \sum_{k=1}^{M} w_n^k \delta(X_n^k, \mathrm{d}x)$.
       Replace $X_n^j \leftarrow Z_n^j$ and $w_n^j \leftarrow 1/M$.
```

Used as is, it leads to the well-known Sampling Importance Resampling (SIR) filter, sometimes also called Bootstrap filter, that can be found in Gordon et al. (1993). A straightforward implementation of the multinomial resampling has complexity $\mathrm{O}(M \log M)$: it is indeed equivalent to simulating M draws from a discrete random variable Z_n such that $\mathbb{P}(Z_n = k) = w_n^k$.

A trivial implementation for such simulations requires first to draw $U_n^1, \ldots, U_n^M$ i.i.d. with uniform distribution and then to find the indexes i_n^j for which $U_n^j \in] \sum_{k=1}^{i-1} w_n^k, \sum_{k=1}^{i} w_n^k]$. Finding the indexes i_n^j has only complexity $\mathrm{O}(M)$ when the random variables are U_n^j are ordered, but ordering these random variables has complexity $\mathrm{O}(M \log M)$ at least, using for instance the quicksort algorithm (see Hoare, 1962).

A practical implementation of the multinomial resampling is proposed in Doucet (1998) which circumvents the naive need of sorting M i.i.d. random variables with uniform distribution and relies upon a direct simulation trick instead. The complexity of the SIR filter can hence be reduced from $\mathrm{O}(M \log M)$ (naive implementation using quicksort) to only $\mathrm{O}(M)$ which saves a significant amount of computational resources.

Residual-multinomial resampling

Residual-multinomial resampling is proposed in Liu and Chen (1998) to reduce the extra variance introduced by the resamping step. It is partially deterministic as opposed to the multinomial resampling and is formulated below. Let $\lfloor x \rfloor$ designate the integer part of a real number x and define for any $n \geqslant 0$:

$$R_n = \sum_{k=1}^{M} \lfloor M \cdot w_n^k \rfloor, \qquad \overline{w}_n^j = \frac{M \cdot w_n^j - \lfloor M \cdot w_n^j \rfloor}{M - R_n}.$$

Algorithm 4.5 (Residual-multinomial resampling step)**.**

```
At time $n \geqslant 0$
    3. Copy $\lfloor M \cdot \widehat{w}_n^j \rfloor$ particles $\widehat{X}_n^j$. ($R_n$ particles are thus allocated, say $Z_n^1, \dots, Z_n^{R_n}$).
       Sample the remaining particles $Z_n^{R_n+1}, \dots, Z_n^M \sim \sum_{k=1}^{M} \overline{w}_n^k \delta(X_n^k, \mathrm{d}x)$.
       Replace $X_n^j \leftarrow Z_n^j$ and $w_n^j \leftarrow 1/M$.
```

The details of residual-multinomial resampling are given in Algorithm 4.5 where only the new step is described. In essence, particles with weights greater than $1/M$ are forced into the new sample, and the rest is allocated at random, depending on the remaining probability mass available. Note that the last part of a residual resampling step is basically a multinomial resampling step on the residual probability mass, hence the name.

It is shown to be computationally cheaper than the multinomial resampling, due to the fact that only a fraction of the M particles are randomly allocated. It does not introduce any bias for the estimation and has the added advantage of having a lower variance than that of the multinomial resampling (see Douc and Cappe, 2005, for the proofs).

Other resampling techniques

Stratified and systematic resampling also offer an alternative to the multinomial resampling scheme (see Kitagawa (1996) and Carpenter et al. (1999) or Chen (2003) for a more general overviews). Systematic resampling appears to be another popular choice in the literature for computational reasons even though its variance is not guaranteed to be smaller than that of the multinomial resampling as stated in Douc and Cappe (2005). A short study of these techniques and a numerical comparison of their performance on an example are offered in Cornebise (2009). Note that residual versions of these techniques also exist, where they are substituted to the multinomial sampling used in the second half of Algorithm 4.5.

Limitations of the resampling procedure

The resampling procedure alleviates the degeneracy problem but also introduces practical and theoretical issues (as mentioned in Doucet et al., 2000, for example). From a practical point of view, resampling very obviously limits the opportunity of parallelisation of the algorithm. From a theoretical point of view, simple convergence results are lost due to the fact that after one resampling step the particles are not independent anymore. Moreover, resampling causes the particles with high importance weights to be statistically selected many times: the algorithm thus suffers from the so-called loss of diversity.

4.3.4 Move step

The loss of diversity among the particles following the resample step is usually addressed in the literature with the introduction of yet another additional move step into the algorithm: the idea behind it is to rejuvenate the diversity after the particles have been resampled.

Using MCMC

Doucet et al. (2001); Gilks and Berzuini (2001) present the so-called Resample-Move algorithm in which an MCMC step is used after resampling. This new step relies upon the use of Markov transition kernels with appropriate invariant distributions. Moving the particles according to such kernels formally guarantees the particles still target the distribution of interest but also give them an additional chance to move towards an interesting region of the state space while increasing the diversity of the sample at the cost of an increased computational burden. Doucet and Johansen (2011) underline the possibility of using even non ergodic MCMC kernels for this purpose and also propose to go a step further and rejuvenate not only the current state but also some of the (immediate) past states with the so-called Block Sampling (the computational cost of which is thus even greater).

Using regularisation

Another approach to deal with the loss of diversity is based upon regularisation techniques. Let us define for $x, x^* \in \mathcal{X} \subset \mathbb{R}^{n_x}$

$$K_h(x, x^*) = h^{-n_x} \cdot (\det \Sigma_n)^{-1/2} \cdot K\left(\Sigma_n^{-1/2} \cdot \frac{x - x^*}{h}\right)$$

where K is usually a smooth symmetric unimodal positive kernel of unit mass (hence a probability measure), h is the bandwidth of the kernel, and Σ_n designates the empirical covariance matrix of the sample (see Silverman, 1986, for the idea of whitening the sample via Σ_n).

Algorithm 4.6 (Regularisation step)**.**

At time $n \geqslant 0$

```
    4. Sample $\epsilon_n^j \sim K(x)$, and set $Z_n^j \leftarrow X_n^j + h \cdot \Sigma_n^{1/2} \cdot \epsilon_n^j$.
       Replace $X_n^j \leftarrow Z_n^j$ and keep $w_n^j \leftarrow w_n^j$.
```

Gordon et al. (1993) originally referred to that step as ″jittering‟ since it adds a small amount of noise to each resampled particle. Note that, when used together with the multinomial resampling scheme described in Algorithm 4.4, the resulting combination of the two steps can be reformulated as described in Algorithm 4.7: it is then equivalent to resampling new particles from the smoothed estimated target distribution (using kernel density K).

Algorithm 4.7 (Alternate formulation for the combination of Algorithms 4.4 and 4.6)**.**

```
At time $n \geqslant 0$
    3+4. Sample $Z_n^j \sim \sum_{k=1}^{M} w_n^k K_h(X_n^k, x)$.
         Replace $X_n^j \leftarrow Z_n^j$ and $w_n^j \leftarrow 1/M$.
```

The choice of both the kernel smoothing density K and the bandwidth h obviously has a big impact on the algorithm. The idea is to resample from a density estimated from the particles at time step n that best approximates the true target density. Picking $K(\cdot) = \delta(\cdot, 0)$ the Dirac mass at the origin turns the regularised SMC filter back into a simple SMC filter. From a general point of view, we would like the estimated density to converge as fast as possible towards the true target density as M goes to $+\infty$, since the number of particles will necessarily be limited by the computational resources.

For the Gaussian kernel (among others), Silverman (1986) shows it is possible to compute the optimal bandwidth to use, i.e. the bandwidth that minimises the variance of the density estimate. Although it could be argued that selecting a proper bandwidth is a difficult task, this optimal bandwidth yields good results in practise and at least provides a rough idea about the scaling of h. As is the case with kernel density estimates, the choice of h directly influences the trade-off made between variance and bias of the estimate: if h is chosen too small, the loss of diversity will still be severe, and if h is chosen too large, the filtered density will roughly be estimated as a single kernel, hence introducing a severe bias into the estimation.

The use of the Epanechnikov kernel, proportional to $1 - \|x\|^2$ on the unit ball of the state space, is recommended in Silverman (1986) because it is asymptotically the most efficient, and Doucet (1998) claims it can be difficult to choose a "good" kernel. However, we advocate the use of the Gaussian kernel whenever possible for computational reasons: simulations from the Gaussian kernel are readily available on most machines and come at a computationally cheaper price than simulations from the Epanechnikov kernel. But the non optimality of Gaussian kernel does not outbalance its ease of use, since the choice of the kernel neither affects the order of the bandwidth nor the rate of convergence as stated in DasGupta (2008).

From a general point of view it is also possible to choose a n_x-dimensional kernel under the form of a product of n_x 1-dimensional (possibly distinct) kernels. Such a choice is preferable when some coordinates of the state are bounded. It allows for easier simulations on these coordinates using dedicated truncated kernels whereas a straightforward accept-reject algorithm could turn out to be highly inefficient (with a low acceptance rate) depending on the boundaries of the state space.

Finally, the regularisation can also be done before resampling thus resulting in the so-called pre-regularised particle filter (pre-PRF) as opposed to the post-regularised particle filter presented here. Theoretical convergence results about these regularised filters are available in Oudjane (2000) and Rossi (2004)

4.3.5 Detection and removal of outliers

In order to deal with the sensitivity of the particle filters to outliers, we propose a new additional rule at the end of step 2 of Algorithm 4.1. Its role is to make sure that outliers do not lead to a fully degenerated situation, that the algorithm would not recover from. The details of it are given in Algorithm 4.8 where only the additional rule is described.

Algorithm 4.8 (Online detection and removal of outliers.)**.**

At time $n \geqslant 0$

If the degeneracy problem is critical, consider the observation y_n as missing (see Algorithm 4.3) and rewind back to step 1.

The rule applies only to situations where the degeneracy of the sample is critical: when the importance density chosen is the prior density, it triggers only when the current observation is not predicted efficiently. In that case, we proceed as if the observation was missing. In practise the degeneracy problem is deemed critical when a criterion such as $\mathrm{ESS}(n) < \epsilon \cdot M$ is met, with $\epsilon > 0$ very small.

Observations that do not agree with the model are thus detected online and ignored to prevent immediate degeneracy. This trick is in a way similar to the one introduced in Hu et al. (2008, 2011) where a resample step is iterated until the likelihood of the current observation with regard to the resampled particles is above a given threshold. While both techniques ensure that the particles do not collapse when an outlier is met, the cost paid is different for each. The alteration proposed in Hu et al. (2008, 2011) can be computationally expensive (with an unbounded runtime) but the observation ends up being taken into account, while our own modification is definitively cheaper (with a guaranteed fixed runtime) but discards the observation at hand when it strongly disagrees with the current state of the model. A significant change of state will still be detected in the long run, because considering the observation y_n as missing automatically implies the variance of the state grows larger (which means that, if it were to be repeated, the outlying observation, would seem more likely at the next time step, with regard to the new state).

4.3.6 Choice of the importance density

As previously stated the particle filters rely on the introduction of an importance density that was chosen of the form given in (4.17) i.e.

$$q^\theta(x_{0:n}|y_{0:n}) = q^\theta(x_0|y_0)\prod_{k=1}^{n} q^\theta(x_k|y_{0:k-1}, y_{0:k}).$$

Choosing carefully the importance density q^θ can help reduce the variance of the importance weights and thus alleviate the degeneracy problem. As the choice is abundantly discussed in the literature, we only selected three representative alternative among the many that are available.

Prior density

A default choice consists of taking $q^\theta(x_0|y_0) = \mu^\theta(x_0)$ and $q^\theta(x_n|x_{0:n-1}, y_{0:n}) = f_n^\theta(x_n|x_{n-1})$, i.e. taking the prior density (4.1) of the model as the importance function. This choice works even with missing data (as it does not depend on y_n) and leads to much simpler calculations for the update of the importance weights as can be seen directly in the formulae given in Algorithm 4.9.

Algorithm 4.9 (Sequential Importance Sampling (SIS) for filtering, using the prior density as the importance density)**.**

At time $n = 0$

1. Sample $X_0^j \sim \mu^\theta(x_0)$.
2. Compute $\widetilde{w}_0^j = g_0^\theta(y_0|X_0^j)$ and set $w_0^j \leftarrow \dfrac{\widetilde{w}_0^j}{\sum_{k=1}^M \widetilde{w}_0^k}$.

At time $n \geqslant 1$

1. Sample $X_n^j \sim f_n^\theta(x_n|X_{n-1}^j)$.
2. Compute $\widetilde{w}_n^j = w_{n-1}^j g_n^\theta(y_n|X_n^j)$ and set $w_n^j \leftarrow \dfrac{\widetilde{w}_n^j}{\sum_{k=1}^M \widetilde{w}_n^k}$.

Note that using the prior density as the importance density makes the algorithm propose new particles in a blind way: the new particles are simulated around the current state, not around the upcoming targeted state. With such a choice of importance density, the algorithm becomes especially sensitive to outliers. An annealed version of the prior distribution is proposed in Chen (2003) to help deal with some situations where prior and likelihood do not agree.

Optimal density

Although popular, the choice of the prior density is not optimal: the optimal choice is given by $q^\theta(x_0|y_0) = \pi^\theta(x_1|y_1)$ and $q^\theta(x_n|x_{0:n-1}, y_{0:n}) = \pi^\theta(x_n|y_n, x_{n-1})$ in the sense that it minimises the variance of the importance weights conditional upon the past states and the past observations as can be seen in Doucet et al. (2000). The idea underlying this choice is to take into account the upcoming observation so that particles are not blind to the upcoming state anymore. Most of the time sampling from these optimal distributions is not an option however, and it is usually recommended to approximate them if possible: for example Pitt and Shephard (1999) propose the so-called Auxiliary Particle Filter, Doucet et al. (2000) use the Extended Kalman filter to derive a Gaussian approximation (relying on a local linearisation of the state space model) and van der Merwe et al. (2001) discuss the use of the Unscented Kalman Filter to obtain such approximations (see Wan and van der Merwe, 2000, for the details about implementing the UKF).

Independent density

Let us mention that it is also possible to use an independent importance density (independent with regard to the states and observations) but it is strongly recommended to avoid such a choice because it ″ignores″ both the current and the upcoming states (see Doucet et al., 2000)

4.3.7 Parameter estimation

Thus far, state estimation was discussed conditionally to the fact that the parameter θ was known. However, θ is often unknown and has to be estimated together with the state of the dynamic model. Kantas et al. (2009) offers a comparative review of the possible choices available for parameter estimation, presenting maximum likelihood and Bayesian parameter estimation in the context of an offline or online procedure. We provide here only a brief overview of the Bayesian parameter estimation and direct the interested reader to the original paper for the complete discussion.

One of the first approach considered in the literature for parameter estimation is to extend the state X_n at time n into a new state $Z_n = (X_n, \theta_n)$ with initial distribution $\mu^{\theta_0}(x_0)\pi(\theta_0)$ and transition density $f^{\theta_n}(x_n|x_{n-1}) \cdot \delta(\theta_n, \theta_{n-1})$ and then estimate this new extended model with a standard SMC filter as in Kitagawa (1996). Even though the approach is theoretically sound as claimed in Kantas et al. (2009); Rossi (2004), it can lead to a strong loss of diversity problem on the coordinate θ when no move step is implemented as the parameter space is only explored at the initialisation of the algorithm, making such an approach often unusable. The addition of a move step into the algorithm provides a satisfying solution to this problem as can be seen in Rossi (2004) who successfully applied the kernel regularisation technique, or in Andrieu et al. (1999) who makes use of MCMC techniques in a move step to update the parameter value. Another option is to force a fictitious small dynamic upon the parameter as described in Higuchi (2001); Kitagawa (1998); Liu and West (2001) so that it is artificially allowed to evolve over time, even though Kantas et al. (2009) rightly remarks that modifying the model in such a way makes it hard to quantify how much bias is introduced in the resulting estimates.

A more recent way of estimating the parameter together with the state relies upon the use of so-called Particle Markov Chain Monte Carlo (PMCMC) methods found in Andrieu et al. (2010). These methods are computationally expensive both in term of storage and calculations, because their computational cost typically grows with time as underlined in Chopin et al. (2012), and thus are less fit for online estimation than some standard SMC filter: the most basic PMCMC method, known as the Particle Marginal Metropolis-Hastings (PMMH) sampler and described in Kantas et al. (2009), involves running an SMC filter for each step of a Metropolis-Hastings algorithm used to propose a new value of the parameter θ.

4.3.8 Asymptotic properties

As the voluminous literature on the topic attests, particle filters are an effective mean of approximating the targeted filtered density. In the recent years, the huge popularity of these methods has drawn the attention of the scientific community upon the theoretical problems underlying their use. Although the convergence (of the approximated filtered distribution towards the true filtered distribution, as the number of particles M goes to $+\infty$) is rather trivial for the SIS filter given in Algorithm 4.1, such results are noticeably harder to get as soon as resample and move steps are involved (the difficulty stemming from the interaction involved within the particles). Some of the most recent and influential works on the matter include Douc and Moulines (2012); Moral and Guionnet (1999); Oudjane and Rubenthaler (2005); Rossi (2004) with Chopin (2004); Crisan and Doucet (2002); Douc and Moulines (2008); Hu et al. (2008) making for a somewhat easier read for the practitioners.

4.3.9 Summary

In the end, keeping in mind that the original aim is the online estimation and prediction, we implemented an algorithm not too computationally expensive. We chose the importance density to be the prior density of the model and included a residual resample step coupled with a Gaussian kernel regularisation move step, that triggered whenever $\mathrm{ESS}(n) < 0.5M$ unless $\mathrm{ESS}(n) < 0.001M$, in which case the current observation was instead considered an outlier and thus treated as missing. As for the parameter estimation problem, we opted for the solution of extending the state-space and introduced no artificial dynamic on the parameter θ, which results in the disappearance of the θ superscript on densities μ, f_n and g_n in the description of Algorithm 4.10. We did however test the introduction of an artificial dynamic on the parameters but observed no changes in the measured overall performance.

Algorithm 4.10 (Particle filter used for our application).

```
At time $n = 0$
    1. Sample $\widehat{X}_0^j \sim \mu(x_0)$.
    2. Compute $\widetilde{w}_0^j = g_0(y_0|X_0^j)$ and set $\widehat{w}_0^j \leftarrow \frac{\widetilde{w}_0^j}{\sum_{k=1}^M \widetilde{w}_0^k}$.
        • if $\widehat{\mathrm{ESS}}(0) < 0.001M$, set $X_0^j \leftarrow \widehat{X}_0^j$ and $w_0^j \leftarrow 1/M$.
        • if $0.001M \leqslant \widehat{\mathrm{ESS}}(0) < 0.5M$, use residual-multinomial resample (see Algorithm 4.5) and
          regularisation move (see Algorithm 4.6) steps to set $X_0^j$ and $w_0^j$.
        • if $0.5M \leqslant \widehat{\mathrm{ESS}}(0)$, set $X_0^j \leftarrow \widehat{X}_0^j$ and $w_0^j \leftarrow \widehat{w}_0^j$.
At time $n \geqslant 1$
    1. Sample $\widehat{X}_n^j \sim f_n(x_n|X_{n-1}^j)$.
    2. Compute $\widetilde{w}_n^j = w_{n-1}^j g_n(y_n|X_n^j)$ and set $\widehat{w}_n^j \leftarrow \frac{\widetilde{w}_n^j}{\sum_{k=1}^M \widetilde{w}_n^k}$.
        • if $\widehat{\mathrm{ESS}}(n) < 0.001M$, set $X_n^j \leftarrow \widehat{X}_n^j$ and $w_n^j \leftarrow w_{n-1}^j$.
        • if $0.001M \leqslant \widehat{\mathrm{ESS}}(n) < 0.5M$, use residual-multinomial resample (see Algorithm 4.5) and
          regularisation move (see Algorithm 4.6) steps to set $X_n^j$ and $w_n^j$.
        • if $0.5M \leqslant \widehat{\mathrm{ESS}}(n)$, set $X_n^j \leftarrow \widehat{X}_n^j$ and $w_n^j \leftarrow \widehat{w}_n^j$.
```

4.4 Application

In this Section we describe an application of particle filters for electricity load forecasting. We quickly describe the data used for our experimentation and the two similar models that were estimated using Algorithm 4.10, deal with the problem of initialising the particle filter and discuss the results obtained.

4.4.1 Data

Calendar information

Time range. The data chosen for the application contain the consolidated half-hourly electricity load at the "EDF" perimeter over the period ranging from 04/01/2006 to 03/31/2011 which represents five years worth of measurements, with 48 points per day. Note that only an estimation of the load is available in real time. The consolidated data correspond to the true (not estimated) signal that is available only three weeks later.

Daytypes. The calendar used for the application provides nine distinct daytypes, the list of which is given in Table 4.1. In essence, this is a very basic calendar that models a single bank-holidays effect where more detailed calendars would model multiple different ones. Although such a basic calendar arguably does not reflect the whole variety of daytypes, it is detailed enough for our purpose and helps keep the dimension of the model we propose as low as possible.

#	day	#	day	#	day
0	mon.	3	sat.	6	BH
1	tue.-wed.-thu.	4	sun.	7	after BH
2	fri.	5	before BH	8	between BH and a weekend

Table 4.1: Daytypes provided by the basic calendar used in the application. BH stands for a bank holiday.

Note that the operational model used by EDF also require the precise specification of daytypes and so-called offsets, the latter being used to model breakpoints (see Bruhns et al., 2005, for the details).

From here on, we will call special, the instants in the calendar where specific information is needed for the operational model to be correctly estimated and predicted. These special instants essentially correspond to bank-holidays (daytypes from 5 to 8), or the summer and winter holiday breaks and are signalled on Figure 4.2.

Temperature information

Two different kinds of temperature related data are available for our experimentation. First of all, the data include the raw observed temperature that we will denote T^{raw} for each instant within the period of study. A so called smoothed heating temperature (because it appears in the heating part of the model) is available over the period of study as well that we will denote T^{heat}. This smoothed heating temperature T^{heat} is a convex combination of the raw temperature T^{raw} and two exponentially smoothed versions of it as described in Bruhns et al. (2005). Let us define $T^{\vartheta}_{n,i}$ an exponentially smoothed version of the raw temperature (with smoothing coefficient ϑ) for the day $n \geqslant 0$ and the instant $0 \leqslant i \leqslant 47$ as

$$\begin{aligned} T^{\vartheta}_{0,0} &= T^{\text{raw}}_{0,0} \\ T^{\vartheta}_{n,i} &= \vartheta \cdot T^{\vartheta}_{n,i-1} + (1-\vartheta) \cdot T^{\text{raw}}_{n,i} \end{aligned}$$

where $T^{\vartheta}_{n,-1}$ is a shorthand for $T^{\vartheta}_{n-1,47}$. Then T^{heat} is defined by

$$T^{\text{heat}}_{n,i} = \alpha_i \cdot T^{\vartheta_{1,i}}_{n,i} + \beta_i \cdot T^{\vartheta_{2,i}}_{n,i} + (1-\alpha_i-\beta_i) \cdot T^{\text{raw}}_{n,i}.$$

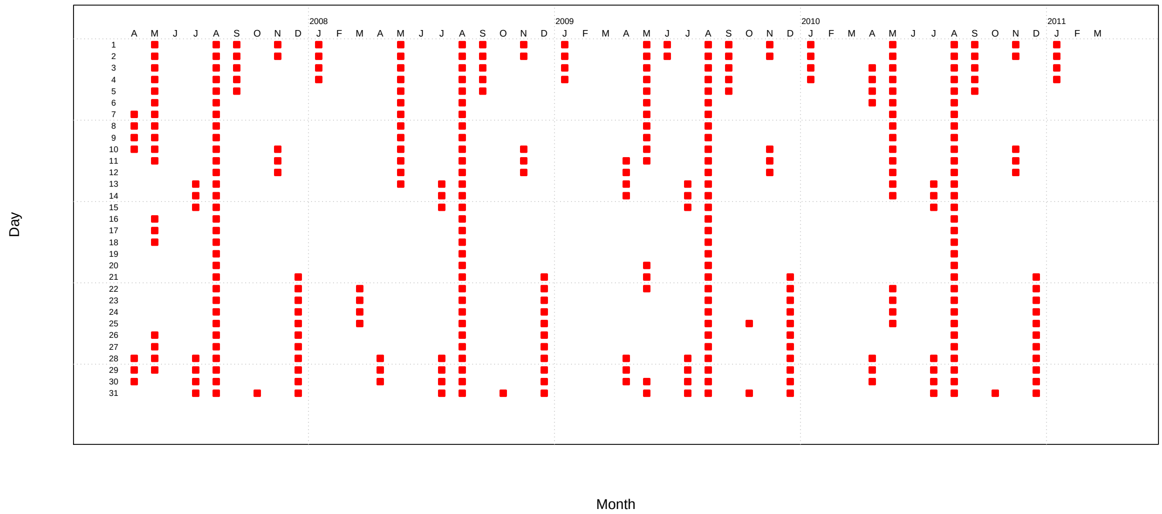


Figure 4.2: Repartition of the bank holidays amidst the calendar from 04/01/2007 to 03/31/2011. Each column represents a month.

where coefficients $\vartheta_{1,i}, \vartheta_{2,i}, \alpha_i, \beta_i$ depend on the instant i considered and are optimised for the heating part of an operational model that was estimated on a part of the data available. Similarly, a smoothed cooling temperature that we will denote T^{cool} is also provided, that only involves a single exponential smoothing instead of two, and the coefficients of which are optimised for the cooling part of the same operational model. Finally from T^{cool}, another quantity Δ^{cool} is built which is defined as

$$\Delta^{\text{cool}}_{n,i} = (T^{\text{cool}}_{n,i} - u^{\text{cool}}_i)\mathbb{1}_{]-\infty, T^{\text{cool}}_{n,i}[}(u^{\text{cool}}_i)$$

and where the coefficients u^{cool}_i were also optimised.

4.4.2 Dynamic models

The formulation of the models that we consider was inspired by the works of Bruhns et al. (2005); Dordonnat (2009); Launay et al. (2012b). It features three parts (seasonality, heating, cooling) similarly to what was done in Launay et al. (2012b) (see Chapter 3) and includes a two layers dynamic on the two most relevant parts with regard to the French electricity load for each of the 48 half-hours (or instants) within a day. The 48 corresponding independent models are estimated and predicted in parallel, using the calendar and temperature information described above, the results being aggregated back together at the end of the process. The dimensions of the parameter and state spaces were voluntarily kept small: the goal is ultimately to provide competitive one-day-ahead predictions for the electricity load based on a model as parsimonious as possible within a rather general framework.

Main model

We denote $\mathcal{N}(\mu, \Sigma)$ the Gaussian distribution with mean μ and variance Σ, and $\mathcal{N}(\mu, \Sigma, \mathcal{S})$ the corresponding truncated Gaussian distribution the support of which is $\mathcal{S}$. For each half-hour, and removing the now superfluous i subscript, the main model that we consider is defined by

$$y_n = x_n + \nu_n, \tag{4.23}$$

where $\nu_n \sim \mathcal{N}(0, \sigma^2)$ and where the state x_n is made of three parts

$$x_n = x_n^{\text{season}} + x_n^{\text{heat}} + x_n^{\text{cool}},$$

that are defined by

$$\begin{aligned} x_n^{\text{season}} &= s_n \cdot \kappa_{\text{daytype}_n} \\ x_n^{\text{heat}} &= g_n^{\text{heat}}(T_n^{\text{heat}} - u^{\text{heat}})\mathbb{1}_{]T_n^{\text{heat}},+\infty[}(u^{\text{heat}}) \\ x_n^{\text{cool}} &= g^{\text{cool}}\Delta_n^{\text{cool}}. \end{aligned}$$

The various components obey the following prior dynamic

$$\begin{aligned} s_n &= s_{n-1} + \epsilon_n^s, & \epsilon_n^s &\sim \mathcal{N}\left(0, \sigma_{s,n}^2,]-s_{n-1}, +\infty[\right) \\ g_n^{\text{heat}} &= g_{n-1}^{\text{heat}} + \epsilon_n^g, & \epsilon_n^g &\sim \mathcal{N}\left(0, \sigma_{g,n}^2,]-\infty, -g_{n-1}^{\text{heat}}[\right) \\ \sigma_{s,n} &= \sigma_{s,n-1} + \eta_n^s, & \eta_n^s &\sim \mathcal{N}\left(0, \sigma_s^2,]-\sigma_{s,n-1}, +\infty[\right) \\ \sigma_{g,n} &= \sigma_{g,n-1} + \eta_n^g, & \eta_n^g &\sim \mathcal{N}\left(0, \sigma_g^2,]-\sigma_{g,n-1}, +\infty[\right) \end{aligned}$$

where daytype_n, T_n^{heat} and Δ_n^{cool} correspond to the exogenous variables that we already discussed:

- denoting N_{daytype} the number of different daytypes featured in the calendar provided, $\text{daytype}_n \in \mathbb{N}$ takes a finite number of values between 0 and $N_{\text{daytype}} - 1$ and represents the class to which the day n belongs with regard to the calendar ;
- $T_n^{\text{heat}} \in \mathbb{R}$ is the temperature used to compute the heating part of the model ;
- $\Delta_n^{\text{cool}} \in \mathbb{R}_+$ provides the cooling degrees needed to compute the cooling part of the model.

Using the definitions and notations introduced in Section 4.1, the parameter of the model is given by $\theta = (\sigma_s, \sigma_g, g^{\text{cool}}, u^{\text{heat}}, \kappa, \sigma)$, these quantities are assumed constant over time in the model. At time n, the state of the model is given by x_n whose components $(s_n, g_n, \sigma_{s,n}, \sigma_{g,n}) \in \mathbb{R}^4$ are the quantities that vary over time according to the dynamic specified. All these quantities are unknown and are to be estimated.

The model (4.23) includes a seasonal part x_n^{season} that is essentially made of a signal s_n, the dynamic prior of which is a random-walk process whose standard deviation $\sigma_{s,n}$ itself evolves as a random-walk. s_n is multiplied by a coefficient $\kappa_{\text{daytype}_n}$ that depends on the daytype of the current observation to model the difference in behaviour between the electricity load on weekdays and weekends or holidays. For identifiability reason, the sum of the coefficients κ_j is fixed so that

$$\frac{1}{N_{\text{daytype}}} \sum_{j=1}^{N_{\text{daytype}}} \kappa_j = 1.$$

Note that s_n essentially replaces the truncated Fourier series featured in Launay et al. (2012b).

The model (4.23) also includes two weather-related parts to account for the influence of low (heating part) and high temperatures (cooling part) upon the electricity load : the heating part x_n^{heat} is based upon a truncated difference between the temperature T_n^{heat} and a heating threshold u^{heat}, as studied in Launay et al. (2012a). This difference is multiplied by a gradient g_n^{heat} whose dynamic is similar to that of s_n: the prior is a random-walk whose standard deviation $\sigma_{g,n}$ itself evolves as a random-walk. Because the cooling effect in France is of a lesser magnitude than the heating effect, the corresponding model for the cooling part x_n^{cool} is simpler: the precomputed truncated difference Δ_n^{cool} is given to the model and multiplied by a cooling gradient g^{cool}.

Notice that to ensure the different quantities involved kept consistent signs throughout time, we specifically used truncated Gaussian distributions. In particular, this means that the random-walks featured in the dynamic are not symmetric and hence that the mean of the state is a priori expected to slightly evolve over time. The constraint on ϵ_n^s and ϵ_n^g can of course easily be lifted if need be, and does not affect the overall predictive performance of the model in any way.

Autonomous model

The dynamic model introduced in (4.23) makes use of the exogenous variables T_n^{heat} and Δ_n^{cool} that were built using coefficients optimised for an operational model, and it thus depends on that model. Since Δ_n^{cool} is only related to the cooling part of the model, and because this is the lest prominent part of the model for the French electricity load, it is usually computed using fixed predefined values for the smoothing and threshold coefficients, and is thus not optimised for any model in truth. The model (4.23) hence depends on that other model only via the exogenous variable T_n^{heat}. To remove this link and obtain a model that can be used on its own, we replace the exogenous variable T_n^{heat} with another temperature variable in the definition of the model.

Let us recall first that T_n^{heat} is built from the raw observed temperature as a mixture between it and two exponentially smoothed versions of it. A straightforward approach to replace the precomputed variable T_n^{heat} is thus to introduce the smoothing coefficients as well as the weights of the mixture as new parameters into the model and estimate the resulting new model. Unfortunately the weights and smoothing coefficients involved in the definition of T_n^{heat} typically vary with the instant of the day and the estimations of these new parameters require the complete time-series of raw observed temperatures (with an half-hour resolution): this means that the amount of calculations involved is massively increased (one exponential smoothing to be computed per particle and per time step), which makes it impossible to estimate such a model online without the corresponding computational resources.

Another straightforward approach is to replace the smoothed heating temperature T_n^{heat} by its original raw observed version T_n^{raw}. Though we do not present the detailed results here, let us mention that the predictions returned by the resulting model were of very poor quality, exceeding a predictive MAPE of 2% for non special instants, even for one-day-ahead forecasts (whereas predictions from the operational model have a MAPE close to 1.2%).

Because estimating the smoothing coefficients online is not an option, we opt for a simpler alternative formulation of the temperature T_n^{raw} as a mixture between the raw observed temperature and an exponentially smoothed version of it

$$T_n = p \cdot T_n^{\vartheta} + (1 - p) \cdot T_n^{\text{raw}}$$

where $p \in [0,\, 1]$ is a new parameter in the model that needs to be estimated and where ϑ is fixed and known. The idea behind this new formulation is to let the model find its own temperature somewhere between the raw observed temperature and a highly smoothed version of it that is still precomputed (thus avoiding a massive increase of computational burden) but not related to any model in particular. As mentioned earlier, we need to fix ϑ : large enough so that the parameter p actually matters, but not too close to the maximal value of 1 either. We choose $\vartheta = 0.98$ which leads to satisfactory results in practise.

For each half-hour, the autonomous model that we consider is hence defined by

$$y_n = x_n + \nu_n, \tag{4.24}$$

where $\nu_n \sim \mathcal{N}(0, \sigma^2)$ and where the state x_n is made of three parts

$$x_n = x_n^{\text{season}} + x_n^{\text{heat}} + x_n^{\text{cool}},$$

that are defined by

$$\begin{aligned} x_n^{\text{season}} &= s_n \cdot \kappa_{\text{daytype}_n} \\ x_n^{\text{heat}} &= g_n^{\text{heat}}(T_n - u^{\text{heat}})\mathbb{1}_{]T_n,+\infty[}(u^{\text{heat}}) \\ x_n^{\text{cool}} &= g^{\text{cool}}\Delta_n^{\text{cool}} \end{aligned}$$

with

$$T_n = p \cdot T_n^{\vartheta} + (1-p) \cdot T_n^{\text{raw}}.$$

The various components again obey the following prior dynamic

$$\begin{aligned} s_n &= s_{n-1} + \epsilon_n^s & \epsilon_n^s, &\sim \mathcal{N}\left(0, \sigma_{s,n}^2,]-s_{n-1}, +\infty[\right) \\ g_n^{\text{heat}} &= g_{n-1}^{\text{heat}} + \epsilon_n^g & \epsilon_n^g, &\sim \mathcal{N}\left(0, \sigma_{g,n}^2,]-\infty, -g_{n-1}^{\text{heat}}[\right) \\ \sigma_{s,n} &= \sigma_{s,n-1} + \eta_n^s & \eta_n^s, &\sim \mathcal{N}\left(0, \sigma_s^2,]-\sigma_{s,n-1}, +\infty[\right) \\ \sigma_{g,n} &= \sigma_{g,n-1} + \eta_n^g & \eta_n^g, &\sim \mathcal{N}\left(0, \sigma_g^2,]-\sigma_{g,n-1}, +\infty[\right) \end{aligned}$$

and where the exogenous variables daytype_n and Δ_n^{cool} are the same as in model (4.23) Notice that such an autonomous model is expected to be somewhat less competitive compared to the main model (4.23) because the heating temperature involves only one exponentially smoothed temperature instead of two. The reason why we include only one such smoothed temperature is that including two, with weights p_1 and p_2, would inevitably lead to identification problems since these temperatures would obviously be highly correlated.

4.4.3 Initialisation of the particle filter

As was already discussed, the degeneracy of the particles sample over time is a serious matter. The choice of the initial distribution of the state is thus of the utmost importance because a strong disagreement between this distribution and the first filtered distribution could lead to sample degeneracy after only a single time step. Two solutions are theoretically viable to choose the initial prior distribution: one may choose either a vague or an informative distribution.

1. on one hand, a vague prior has the advantage of not biasing the dynamic model before the first observations. However, the variance of the initial distribution of the state being very large, the sequence of posterior variances of the filtered distributions of the state tend to decrease very quickly at first. From an SMC filter point of view, one has to use a very large sample of particles to cover at the same time the regions of the state space with prior highest probability and with posterior highest probability: a vague initialisation thus requires the use of a massive number of particles.

2. on the other hand, designing an informative distribution is a totally different task, but still not a trivial one: one has to keep in mind that a "bad" choice of initial distribution may lead to immediate degeneracy. Intuitively, the ideal solution would be to dispose at time $n = -1$ of a filtered distribution $\pi^{\theta}(x_{-1}|y_{-1}, \ldots, y_{-m})$ to use it as a the initial distribution at time $n = 0$. Such a choice is of course not possible because observations are only available for time $n = 0, \ldots, N$.

Note that the trick of ignoring outliers introduced into the particle filter (see Algorithm 4.8) does not alleviate the problem of initialisation, since it can only increase the variance if it is used.

We thus opted for a more general procedure that allows for an automated initialisation of the particles sample to a fitting state space region from time $n = n_0$, and that combines the two approaches mentioned above to retrieve the benefits of both:

1. we use a vague distribution to estimate the smoothed distribution up to time $n = n_0 - 1$ using open-source MCMC generic software such as BUGS (Lunn et al., 2000) or JAGS (Plummer, 2003): we typically chose $n_0 = 365$ so that the variance of the filtered distribution at time $n = n_0 - 1$ is already small enough not to require the use of a massive amount of particles ;

2. after this first MCMC initialisation phase, we retrieve particles (approximately) distributed along the filtered distribution of the state at time $n = n_0 - 1$: this distribution is the one used (through these particles) to initialise the SMC filter at time n_0.

There is however a price to pay for solving the initialisation problem in such a way. First we have to use MCMC to initialise the particle filter and second it makes it hard to use the particle filter on a time series with few observations. Note that the first issue raised is but rhetorical: MCMC, even if expensive, has to be run only once, and not at each time step.

Initial distribution for the MCMC estimation

The initial distribution envisioned for the main model (4.23) is vague and specified by:

$$\begin{aligned}
s_0, g^{\text{cool}} &\sim \mathcal{N}(0, 10^8, \mathbb{R}_+) \\
g_0^{\text{heat}} &\sim \mathcal{N}(0, 10^8, \mathbb{R}_-) \\
u^{\text{heat}} &\sim \mathcal{N}(14, 1) \\
\kappa / N_{\text{daytype}} &\sim \mathcal{D}_{N_{\text{daytype}}}(1, \ldots, 1) \\
\sigma^2, \sigma_{s,0}^2, \sigma_{g,0}^2, \sigma_s^2, \sigma_g^2 &\sim \mathcal{IG}(10^{-2}, 10^{-2})
\end{aligned}$$

where $\mathcal{D}_d(\alpha_1, \ldots, \alpha_d)$ is the Dirichlet distribution in $\mathbb{R}_+^d$ with parameter α (in particular $\mathcal{D}_d(1, \ldots, 1)$ is the uniform distribution over the simplex of $\mathbb{R}_+^d$ defined by $\sum_{i=1}^d x_i = 1$). For the autonomous model

(4.24) the initial distribution is completed with

$$p \sim \mathcal{U}[0, 1].$$

Practical issue

We faced some technical issues running the MCMC estimation up until time $n_0 = 365$ since the Markov Chain outputs were not usable: even with a large burn-in period, the sample returned would not pass the diagnostic tests for the convergence of the empirical distribution towards the true target (see Gelman and Rubin, 1992, for example). For the initialisation via MCMC we thus separated the initial distribution into two parts, essentially isolating the dynamic on the variance of the random-walks, and proceeded as follows.

First we estimated the models as defined in (4.23) and (4.24) up until time $n_0 = 365$, using MCMC generic software such as described in Lunn et al. (2000); Plummer (2003), with the following modification

$$\sigma_{s,n} = \sigma_{s,n-1} = \sigma_{s,*}$$
$$\sigma_{g,n} = \sigma_{g,n-1} = \sigma_{g,*}$$

with initialisation

$$\sigma^2_{s,*}, \sigma^2_{g,*} \sim \mathcal{IG}(10^{-2}, 10^{-2}),$$

in essence removing the second layer in the dynamic from the models (since $\sigma_{s,n}$ and $\sigma_{s,n}$ are not allowed to vary with time anymore). This led to a posterior distribution on a diminished state, that we denote $\widetilde{\pi}_1(\widetilde{x}_{n_0-1}|y_{0:n_0-1})$. From there we completed this posterior distribution with an additional prior $\widetilde{\pi}_2$ on σ_s and σ_g to serve as an initialisation at for the full models at time n_0.

The initial distributions of the particle filter for the full models at time n_0 were thus of the form

$$\pi(x_{n_0-1}|y_{0:n_0-1}) \propto \widetilde{\pi}_1(\widetilde{x}_{n_0-1}|y_{0:n_0-1}) \times \widetilde{\pi}_2(\sigma_{s,n_0-1}, \sigma_{g,n_0-1})$$

with

$$\sigma^2_s \sim \mathcal{N}(\overline{m}_s, \overline{s}^2_s, \mathbb{R}^*_+)$$
$$\sigma^2_g \sim \mathcal{N}(\overline{m}_g, \overline{s}^2_g, \mathbb{R}^*_+)$$

where $\overline{m}_s, \overline{m}_g, \overline{s}^2_s, \overline{s}^2_g$ were values chosen empirically based on $\widetilde{\pi}_1$: for example, we chose m_s and m_g to be the standard deviations of the posterior MCMC estimated samples $(\mathbb{E}[\epsilon^s_1|y_{0:n_0}], \dots, \mathbb{E}[\epsilon^s_{n_0}|y_{0:n_0}])$ and $(\mathbb{E}[\epsilon^g_1|y_{0:n_0}], \dots, \mathbb{E}[\epsilon^g_{n_0}|y_{0:n_0}])$ respectively.

4.4.4 Predictions

Quality criterion

To assess the quality of the models we propose, we will mainly look at their respective predictive performances measured by Mean Absolute Percentage Error (MAPE). As a matter of fact, we are working with half-hourly data and we will model each half-hour independently from one another, a common choice given the type of data, thus leading to 48 separate daily model (see Section 4.4.2). Indexing the respective MAPE criteria of these models by the instant $i = 0, \dots, 47$ to which they are associated, and given their respective observations $y_{1,i}, \ldots, y_{n,i}$, these models return 48 τ-day-ahead predictions defined as the expectations of the predictive distributions i.e. for $i = 0, \ldots, 47$

$$\widehat{y}_{n+\tau,i} = \mathbb{E}[x_{n+\tau} | y_{0:n,i}]. \tag{4.25}$$

The corresponding predictive (with prediction horizon τ) MAPE criterion that we consider for these 48 models is defined, for $i = 0, \ldots, 47$, by

$$\mathrm{MAPE}_i(\tau) = \frac{1}{n-\tau} \sum_{k=1}^{n-\tau} \left| \frac{\widehat{y}_{k+\tau,i} - y_{k+\tau,i}}{y_{k+\tau,i}} \right|$$

and we will most often aggregate the results as

$$\begin{aligned} \mathrm{MAPE}(\tau) &= \frac{1}{48} \sum_{i=0}^{47} \mathrm{MAPE}_i(\tau) \\ &= \frac{1}{48(n-\tau)} \sum_{i=0}^{47} \sum_{k=1}^{n-\tau} \left| \frac{\widehat{y}_{k+\tau,i} - y_{k+\tau,i}}{y_{k+\tau,i}} \right| . \end{aligned}$$

Operational predictions

We will also compare these models to the so-called operational prediction (available from 01/01/09 only, i.e. for the second half of our dataset only) i.e. the final prediction that was actually used by EDF. Note that the operational prediction $\mathrm{Pred}_{\mathrm{OP}}$ cannot be written as a prediction coming from a statistical model (even though we will sometimes abusively refer to it as the prediction from the operational model) : it combines manual adjustments and statistical models. $\mathrm{Pred}_{\mathrm{OP}}$ is computed as a 50%–50% mixture between the two predictions $\mathrm{Pred}_{\mathrm{DOAAT}}$ and $\mathrm{Pred}_{\mathrm{DCo}}$ that we briefly describe below.

The prediction $\mathrm{Pred}_{\mathrm{DOAAT}}$ is obtained as follows. A model similar to the one described in Bruhns et al. (2005), with an ARIMA part, is first used on a real-time estimated signal corresponding to the ″France″ perimeter. An estimated loss is then substracted from it, accounting for the customers within this perimeter that are not affiliated with EDF. A manual adjustment is finally applied in real-time. It is a “top-down” prediction in the sense that the “EDF“ perimeter is approximated as a difference between the “France“ perimeter and a ″France but not EDF″ perimeter.

The prediction $\mathrm{Pred}_{\mathrm{DCo}}$ is obtained as follows. Multiple models from Bruhns et al. (2005) are used upon consolidated signals (not available in real-time, only three weeks later) for sub-perimeters, the reunion of which is the “EDF“ perimeter. The corresponding predictions are then added together before a manual adjustment is finally applied in real-time. It is a ″bottom-up“ prediction in the sense that the ″EDF“ perimeter is approximated as the sum of all its parts.

There are a number of differences between the dynamic predictions and the operational predictions. First of all, the operational predictions are computed using predicted temperatures (since the sequence of observed temperatures at the time of prediction is clearly not available) whereas the models that we consider (see Section 4.4.2) are based on the realised temperatures. The operational predictions make use of a calendar that includes more daytypes and also benefit from high level expertise through the manual adjustments mentioned. But the biggest difference in nature between these predictions lies somewhere else: the dynamic predictions are made from one day to the next (with no intraday correction whatsoever, since we are basically considering 48 independent models), while the operational predictions are made from one half-hour to the next. Essentially the horizon of prediction for the dynamic models is $\tau = 1$ day $= 48$ half-hours whereas it is much smaller for $\text{Pred}_{\text{DOAAT}}$, since the new data get incorporated approximately every 8 half-hours (the computation of $\text{Pred}_{\text{DOAAT}}$ is based upon a real-time signal though, not consolidated data).

4.4.5 Results

Running the filter

For the estimation and prediction of the models, we used the Algorithm 4.10 with a total number of $M = 10^5$ particles. One time step (filtering and predicting the state with horizon $\tau = 1$, including 90% credible intervals) took approximately 1 second on a single core Intel(R) Xeon(R) E5410 (2.33GHz) for one of the 48 independent models, which is compatible with the goal of being able to predict the electricity load in an online manner. The execution time grew a bit larger and reached 3 seconds per iteration when the predictive horizon was set to $\tau = 5$. Note that providing credible intervals requires the use of a sorting algorithm, for example quicksort (see Hoare, 1962) with complexity $\mathrm{O}(M \log M)$ whereas Algorithm 4.10 has only complexity $\mathrm{O}(M)$. Quicker runtimes are thus obviously achievable if the computation of credible intervals is not needed.

Degeneracy

Before looking at the filtered or predicted distributions that we are most interested in, we actually have to assess whether the numerical results obtained are actually usable or not. If the degeneracy problem proved too strong along the estimation process, the estimated values indeed become questionable.

Figures 4.3, 4.4 and 4.5 show the evolution of the various criteria discussed in Section 4.3.2 throughout time for the main model (4.23) at the instant 12:00. These criteria exhibit a seasonal behaviour (with a 1 year period), as the time series itself, showing that the particle filter is subject to a little more degeneracy during winter than during summer (the electricity load is indeed harder to predict, due to the influence of the temperature). Let us mention that the criteria looked very much the same for the autonomous model (4.24) which is why the corresponding graphics are not included here. Although the coefficient of variation $\mathrm{CV}(n)$ is only a rescaling of the effective sample size $\mathrm{ESS}(n)$ (see (4.21)), the outliers detected by the Algorithm 4.10 used are much easier to spot on Figure 4.5 than on Figure 4.3. Also observe that even if the entropy and the coefficient of variation approximate two different divergences (see Cornebise, 2009, for the details), the outliers are as easily spotted on Figures 4.4 and 4.5 and the behaviours of the two criteria are very similar : hence, using the entropy instead of the effective sample size (or the coefficient of variation, since they are interchangeable) to detect outliers in Algorithm 4.10 could be doable (after having developed a basic intuition of its scaling, in order to decide of a threshold) but would not change the results obtained in any major way.

Outliers

We show in Figure 4.6 the number of data that were automatically detected as outliers by both models for each instant (half-hour) of the day. Recall that, according to Algorithm 4.10, an outlier is detected whenever the effective sample size would have dropped below 0.1% of the actual sample size. The amount of outliers varies from one half-hour to the next because an observation flagged as an outlier at a given instant does not necessarily imply that the observation at the next instant will also be flagged. In particular, we observe that more outliers are detected during the day than during the night, which suggests that nighttime is slightly easier to predict than daytime (recall that outliers are essentially data

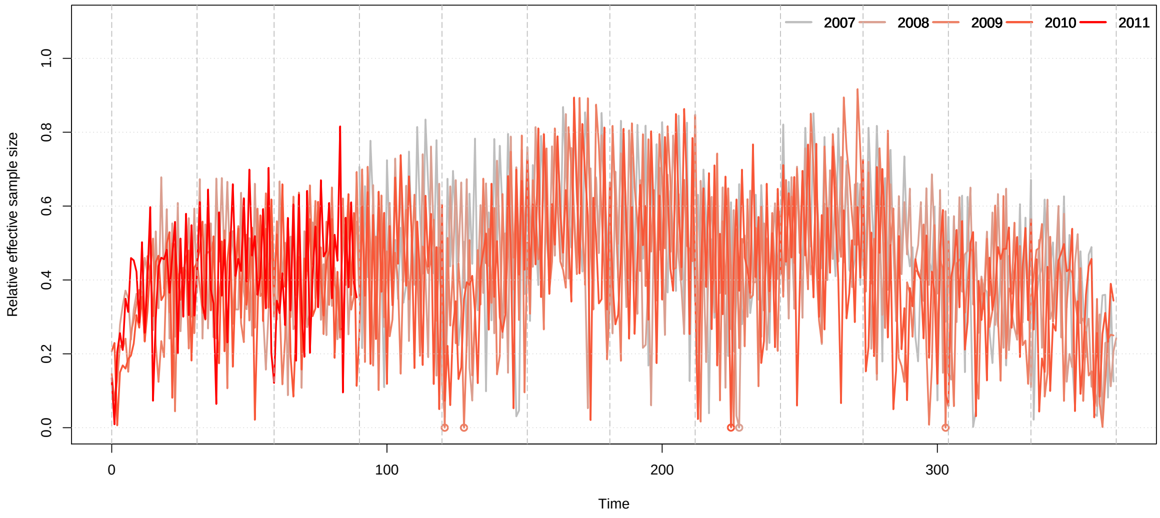


Figure 4.3: Relative effective sample size $\frac{\mathrm{ESS}(n)}{M}$ for the main model (4.23) at 12:00 as a function of the day in the calendar. The saturation of the colour used increases with each year. Data detected as outliers are marked with a circle.

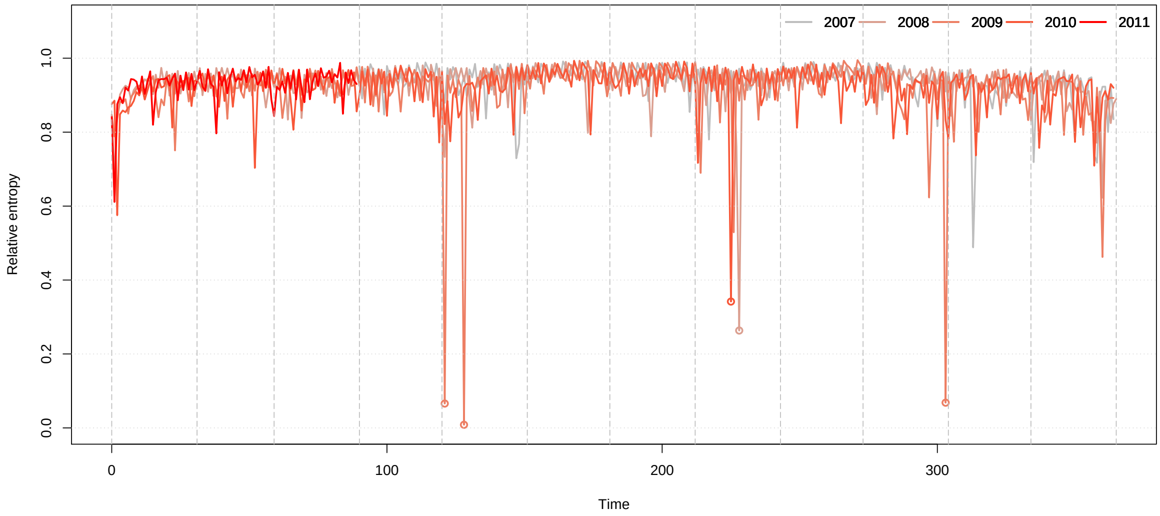


Figure 4.4: Relative entropy $\frac{\mathcal{E}(n)}{-\log M}$ for the main model (4.23) at 12:00 as a function of the day in the calendar. The saturation of the colour used increases with each year. Data detected as outliers are marked with a circle.

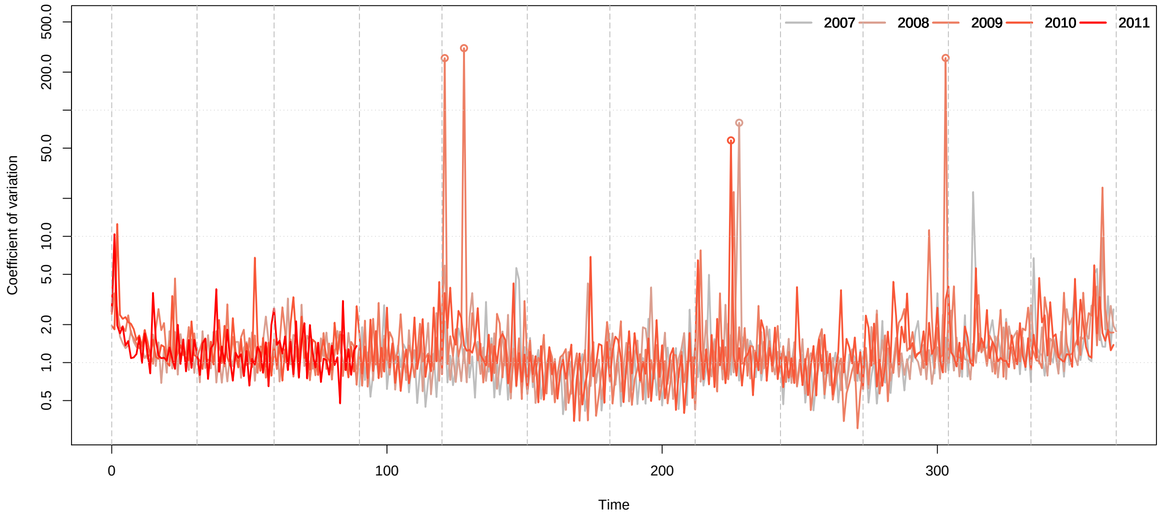


Figure 4.5: Coefficient of variation CV(n) for the main model (4.23) at 12:00 as a function of the day in the calendar. The saturation of the colour used increases with each year. Data detected as outliers are marked with a circle. The ordinate axis is in log-scale.

that are badly predicted). As can be seen on the Figures shown, the amount of variation from one model to the other is rather small, as far as the number of outliers per instant is concerned (but that is only logical, recalling that they exhibited similar degeneracy criteria throughout time).

Figure 4.7 shows the number of outliers depending on the calendar. It allows us to pinpoint the times of the year at which these outliers are actually detected. The summer and winter holiday breaks, and the daylight saving time adjustments are easily spotted. Note that for these events, no prior information was available to the dynamic models. Some days before or after bank holidays are also flagged as outliers (05/02, 05/02, 11/10), even though the dynamic model benefits from some calendar information. This should not come as a surprise however: the daytype specification that we chose is rather poor compared to the calendar used for the operational predictions. A more refined calendar, involving specific daytypes, is likely to help turning these few outliers back into regular data, provided the initialisation of the particle filter is correctly done.

Table 4.2 summarises what is already guessable from Figures 4.2 and 4.7, i.e. that most of the (few) instants detected as outliers by the dynamic models are indeed special instants (recall that special instants are instants in the calendar where specific information is needed for the operational model to be correctly estimated and predicted).

Performance and instants

We show the overall predictive (horizon $\tau = 1$) performance of the dynamic models (4.23) and (4.24) against the operational model (OP) in Table 4.3, depending on whether special instants were included in the calculations or not. The results shown in both cases aggregate the 48 models that were estimated independently from one another. Over the whole period of study, the operational predictions are better

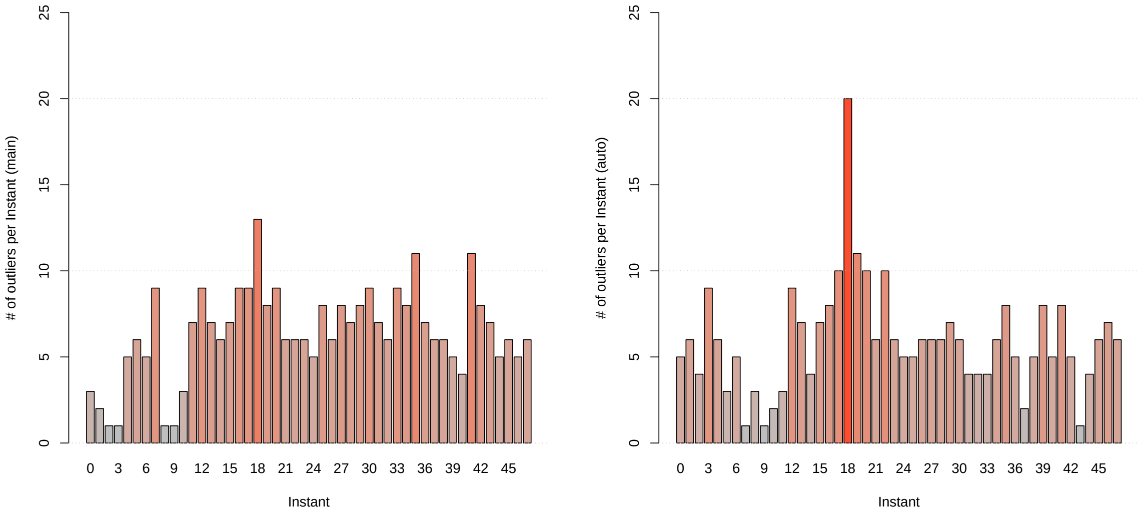


Figure 4.6: Number of outliers detected for each instant of the day by the main model (4.23) (left) and by the autonomous model (4.24) (right).

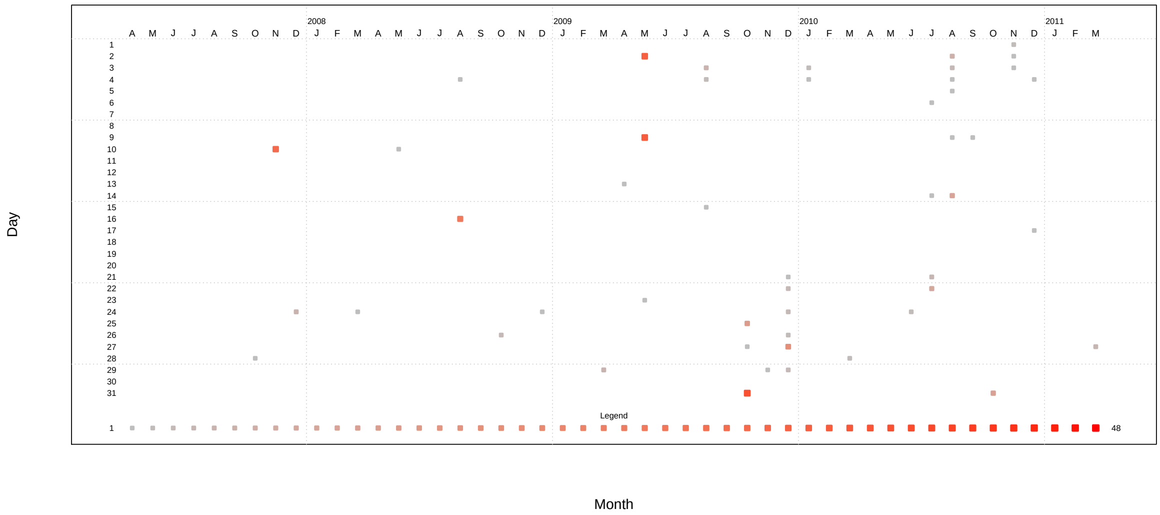


Figure 4.7: Number of outliers detected by the main model (4.23) depending on the calendar from 04/01/2007 to 03/31/2011. Each column represents a month. The size of the point and the saturation of the colour used grow with the number of outliers, as indicated in the legend beneath.

instant	outlier	not outlier
special	269 [5.60]	16627 [346.40]
not special	38 [0.79]	53194 [1108.21]

instant	outlier	not outlier
special	238 [4.96]	16658 [347.04]
not special	47 [0.98]	53185 [1108.02]

Table 4.2: Classification of the instants for the main model (4.23) (left) and for the autonomous model (4.24) (right). The number given between square brackets is an equivalent of the number of instants in days (i.e. divided by 48).

than the predictions provided by the dynamic models, but they also do benefit from more specific calendar information being used to compute them. When only the non special instants are considered, i.e. when holiday breaks and bank-holidays are removed from the calculations, the overall predictive quality of the dynamic models improves considerably as demonstrated by the results in Table 4.3.

	main (4.23)	auto (4.24)	OP
overall	1.4342	1.5416	1.2344
no special	1.1712	1.2971	1.2185

Table 4.3: Overall predictive (horizon $\tau = 1$) and MAPE (in %) for the dynamic models and the operational model (OP). The top row results include special instants in the calculations, while the bottom row results do not.

In fact, looking at Figure 4.8 that represents the predictive MAPE of the main model (4.23) and operational model (OP) averaged by instant, we are able to see that the main model (4.23) predicts the electricity load quite well on non special instants, challenging the operational model throughout the day, except during the morning ascent. The good predictive performance of the dynamic models on non special days is somewhat surprising because the dynamic predictions, coming from 48 independent models, are made from one day to the next whereas the operational predictions include an ARIMA adjustment phase to take advantage of the most recent observations, and also benefit from manual adjustments.

Figure 4.9 shows a comparison of the predictive MAPE between the main model (4.23) and the autonomous model (4.24) : the autonomous model yields less accurate predictions than the main model during daytime, i.e between 08:00AM and 08:00PM roughly, while the performances of these two models remain sensibly equivalent during nighttime. Such findings are logical, considering that the link between the temperature and the electricity load is expected to be much more complex and more important during daytime and also considering that nighttime seasonality is expected to be smoother than daytime seasonality. Even so, the autonomous model (4.24) provides reasonable results, taking into account that it is completely independent from any other operational model (recall that model (4.24) does not use the precomputed optimised smoothed heating temperature that the main model (4.23) does).

Performance and daytypes

We show the predictive MAPE, averaged by daytype, for both dynamic models and the operational model on Figure 4.10 again including or excluding special instants. As we previously underlined, the dynamic model does not provide good predictions for special instants (mainly non regular daytypes, i.e.

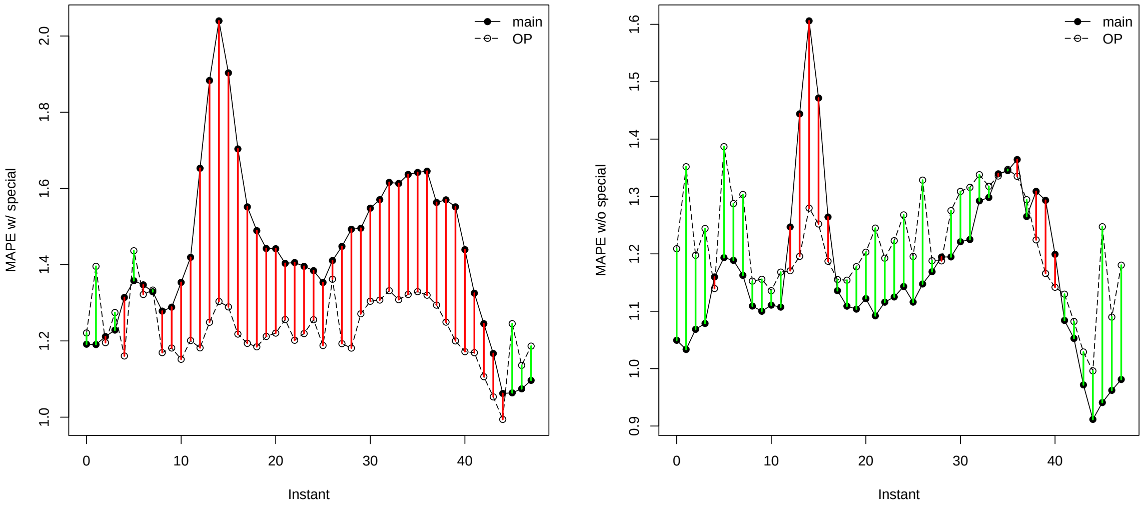


Figure 4.8: Predictive (horizon $\tau = 1$) and MAPE (in %) for the main model (4.23) and the operational model (OP) for each of the 48 half-hours, including special instants in the calculations (leftmost figure) and not including special instants in the calculations (rightmost figure). The difference between the two models is coloured depending on its sign: green when the main model is better than the operational model and red when not.

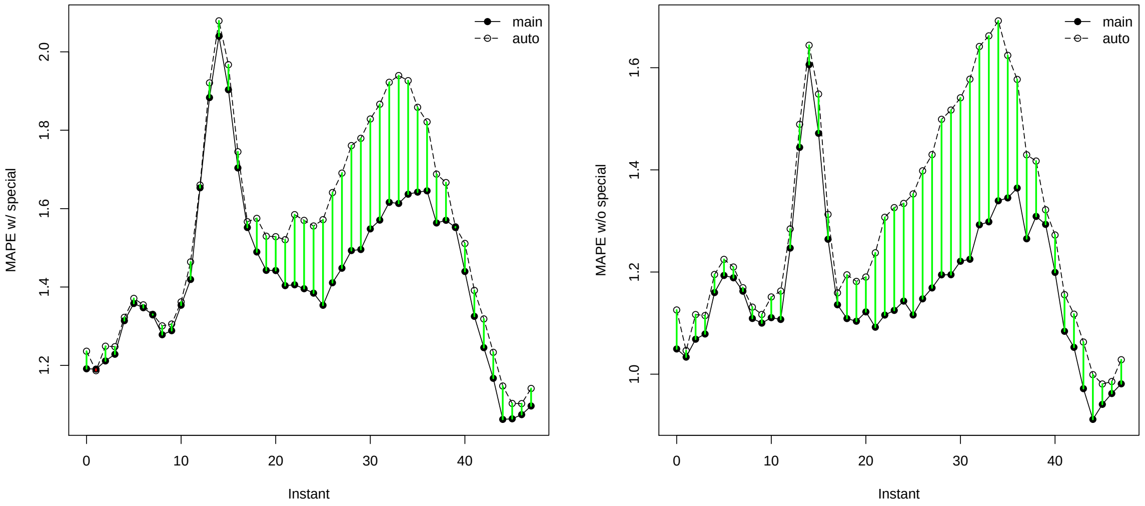


Figure 4.9: Comparison of the predictive (horizon $\tau = 1$) MAPE (in %) of the main model (4.23) and the autonomous model (4.24) when non special instants are included in the calculations (left) and when they are not (right). The difference between the two models is coloured depending on its sign: green when the main model is better than the operational model and red when not.

daytypes 5 to 8). The predictive MAPE, when considering non regular daytypes exclusively, reached 3.34% for the main model (4.23) and 3.46% for autonomous model (4.24).

The relatively poor overall performances of both dynamic models for Mondays (daytype 0) and Sundays (daytype 4) compared to the operational model are also due to the presence of special instants within these daytype classes (typically the summer and winter holiday breaks): once these are removed from the calculations, the dynamic models are seen to remain competitive with the operational model.

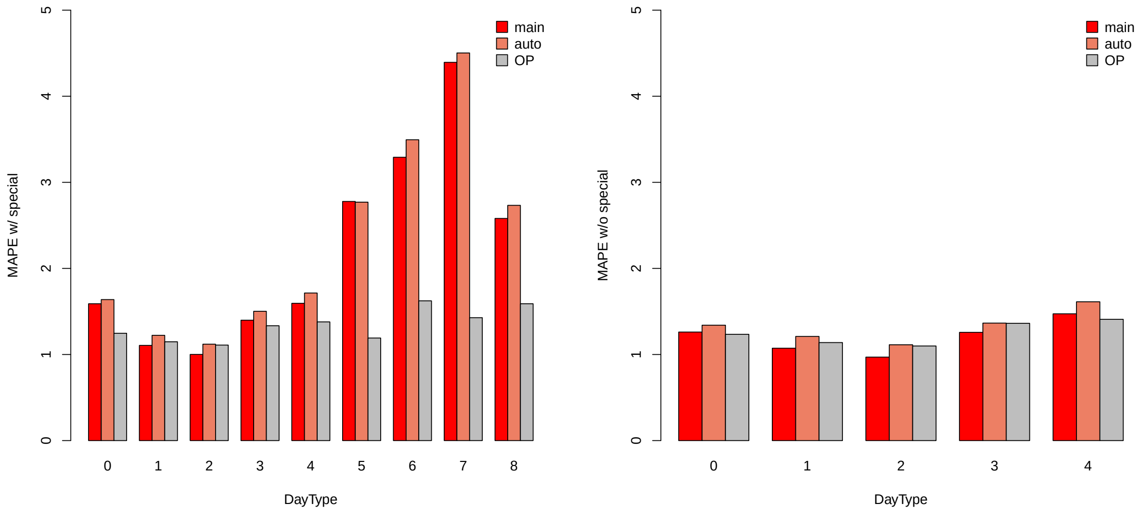


Figure 4.10: Predictive (horizon $\tau = 1$) MAPE (in %) for the dynamic models and the operational model (OP) for each daytype defined in Table 4.1, including special instants in the calculations (left) and not including special instants in the calculations (right).

From here on because the main model (4.23) and the autonomous model (4.24) actually share the same behaviour, we only include Figures and Tables that relates to the main model (4.23).

Performance and temperature

We now focus on the predictive quality of the main model (4.23) depending on the temperature (exogenous variable). Plotting the predictive error of the model against the temperature (be it the raw temperature observed or the smoothed temperature used in the model) does not reveal any specific bias, indicating an overall satisfactory behaviour. As a matter of fact, with the exception of the detected outliers, the predictive error of the main model (4.23) behaves similarly to the predictive error of the operational model as can be seen on Figure 4.11. The dynamic model does not exhibit any visible temperature-based bias but the predictive errors for cold and warm temperatures show slightly more variability than their counterparts for the operational model.

To investigate the reaction of the main model (4.23) to brusque changes in temperature, we show in Figures 4.12 and 4.13 the predictive errors of the main model (4.23) and the operational model against the difference between raw and smoothed temperature. For the dynamic model, a small bias is visible when only the instants for which the raw temperature was colder than 5°C are considered (see Figure

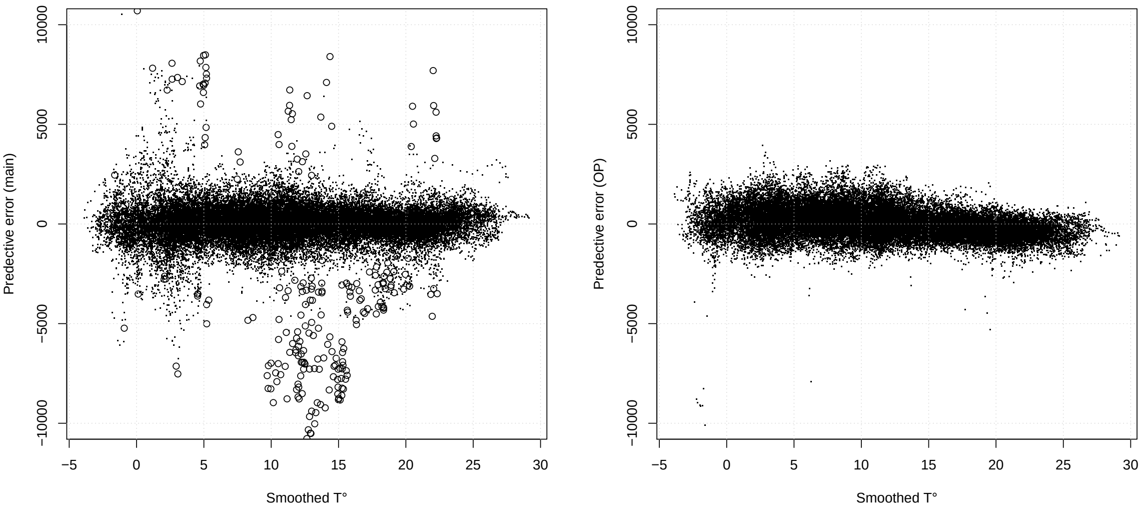


Figure 4.11: Predictive (horizon $\tau = 1$) errors of the main model (4.23) (left) and the operational model (OP) (right) depending on the smoothed temperature. Predictions for data detected as outliers are marked with a circle for the dynamic model.

4.12): it shows us that the dynamic model slightly underestimates (respectively overestimates) the load when the smoothed temperature given to the model is late on the raw temperature and the raw temperature itself is decreasing (respectively increasing), and hints at the possible need of considering a dynamic smoothing model. However no such phenomenon is detected for the dynamic model on the other side of the temperature scale, when looking at raw temperatures warmer than 18°C (see Figure 4.13). Note that the operational model overestimates the electricity load for warm temperatures (see also Figure 4.11) while the dynamic model does not.

Performance and horizon

Since the operational predictions are sometimes required up to $\tau = 3$ days, we now investigate the predictive quality of our dynamic models as the horizon for prediction grows larger. Figure 4.14, given hereafter, displays the predictive MAPE for horizon $\tau = 1, \ldots, 5$, whether including special instants in the calculations or not. It is clear that the predictive errors of the models (4.23) and (4.24) increase with the horizon τ considered for the prediction, confirming that the dynamic model is primarily meant for short-term forecasts and not long-term forecasts.

Another consequence of increasing the prediction's horizon is that the credible intervals obtained around the predictions also tend to grow larger on average as can be observed in Table 4.4. Note that the uncertainty about the temperature introduced in the autonomous model via the parameter p leads to larger credible intervals compared to those of the main model. An illustration of the credible intervals returned by the dynamic models is given in Figure 4.15 where the electricity load is predicted over 48 consecutive instants via the main model (4.23). The predictions clearly improve over time as the model takes more and more recent information into account: the one-day-ahead predictions about 12/30/2010

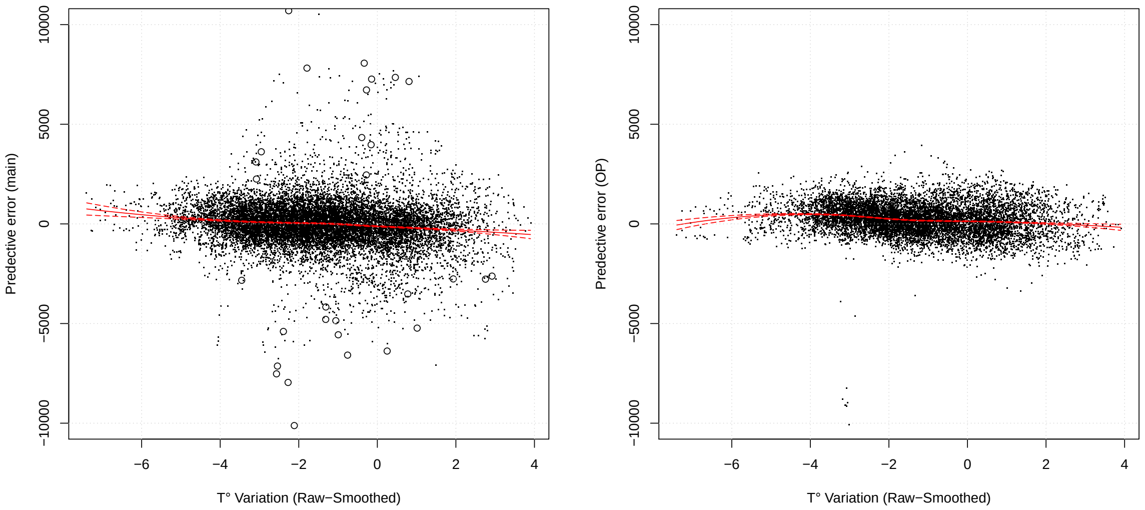


Figure 4.12: Predictive (horizon $\tau = 1$) errors of the main model (4.23) (left) and the operational model (OP) (right) depending on the difference between raw and smoothed temperatures when raw temperature is colder than 5°C. Predictions for data detected as outliers are marked with a circle for the dynamic model. The red line corresponds to an estimation of the bias via the loess function in R with its 90% confidence interval.

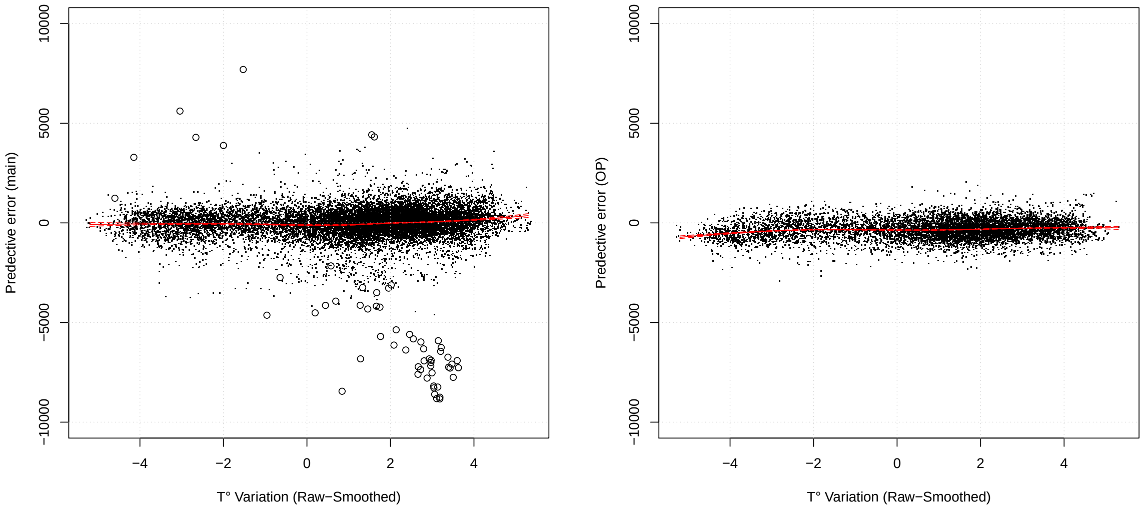


Figure 4.13: Predictive (horizon $\tau = 1$) errors of the main model (4.23) (left) and the operational model (OP) (right) depending on the difference between raw and smoothed temperatures when raw temperature is warmer than 18°C. Predictions for data detected as outliers are marked with a circle for the dynamic model. The red line corresponds to an estimation of the bias via the loess function in R with its 90% confidence interval.

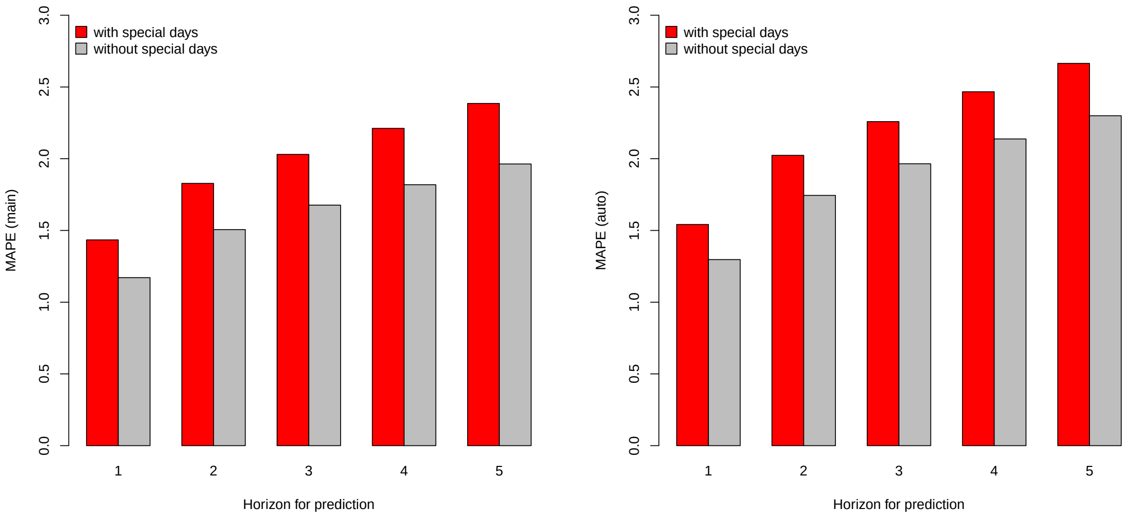


Figure 4.14: Predictive MAPE of the main model (4.23) (left) and the autonomous model (4.24) (right) for $\tau = 1, \ldots, 5$, including special instants in the calculations (leftmost bars) and not including special instants in the calculations (rightmost bars).

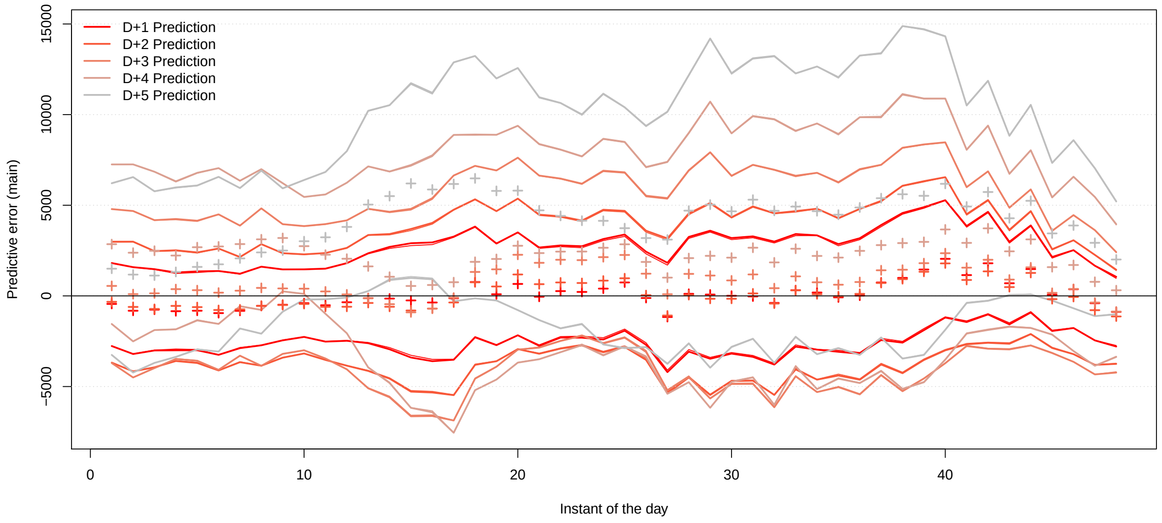


Figure 4.15: Predictive errors (predictive mean minus true value) of the main model (4.23) for the observations on 12/30/2010 (48 half-hours) with $\tau = 1, \ldots, 5$. The horizontal black line marks the true load (no predictive error), and crosses mark the various predictive errors with their respective credible intervals in solid lines. The more recent the predictions are (i.e. the smaller τ is), the more saturated the colour used is: D+1 Prediction is the most recent prediction (it was made 1 day before) while D+5 is the oldest prediction (it was made 5 days before).

provided on 12/29/2010 are much more accurate than the five-days-ahead predictions (of the same day) that were computed on 12/25/2010. Figure 4.15 also makes it clear that the credible intervals obtained for a predictive horizon $\tau = 1$ are narrower compared to those obtained for a predictive horizon $\tau = 5$ (but note that their lengths vary over time in both cases).

main (4.23)	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
$\widehat{\lambda}_{90\%}(x_{n+\tau})$	2746.3	3721.1	4505.6	5191.6	5815.3
$\widehat{\lambda}_{90\%}(y_{n+\tau})$	3036.1	3947.7	4696.5	5358.6	5964.9

auto (4.24)	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
$\widehat{\lambda}_{90\%}(x_{n+\tau})$	3078.3	4216.3	5117.1	5894.7	6597.7
$\widehat{\lambda}_{90\%}(y_{n+\tau})$	3309.9	4394.0	5265.5	6024.1	6713.4

Table 4.4: Mean length (in MW) of the symmetric 90% credible intervals (CI) around the predicted states $\widehat{x}_{n+\tau}$ and around the predicted observations $\widehat{y}_{n+\tau}$ of the main model (4.23) (top) and the autonomous model (4.24) (bottom), for $\tau = 1, \ldots, 5$.

main (4.23)	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
$\widehat{\chi}_{90\%}(\widehat{x}_{n+\tau})$	89.569	90.442	92.385	93.479	94.168
$\widehat{\chi}_{90\%}(\widehat{y}_{n+\tau})$	92.531	92.501	93.773	94.472	94.882

auto (4.24)	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
$\widehat{\chi}_{90\%}(\widehat{x}_{n+\tau})$	90.566	90.905	92.790	93.863	94.465
$\widehat{\chi}_{90\%}(\widehat{y}_{n+\tau})$	92.785	92.394	93.743	94.492	94.958

Table 4.5: Empirical coverage (in %) of the symmetric 90% credible intervals (CI) around the predicted states $\widehat{x}_{n+\tau}$ and around the predicted observations $\widehat{y}_{n+\tau}$ of the main model (4.23) (top) and the autonomous model (4.24) (bottom), for $\tau = 1, \ldots, 5$.

The empirical coverages of the symmetric 90% credible intervals around the predicted states and observations are given in Table 4.5. These values were computed as the ratio between the number of instants for which the observations fell inside the interval, and the total number of instants. Note that if the observations were mutually independent outcomes of the same random variable (which they are not in our situation because of the exogenous variables temperature and calendar), this ratio would theoretically approximate the true rate of coverage i.e. 90%. Even so, the empirical coverage computed seems, somewhat reassuringly, to agree with the expected rate.

Filtered weather parts

Figures 4.16 and 4.17 show the filtered heating and cooling parts of the main model (4.23). Both of these parts actually seem to be piecewise linear with regard to the temperature variables upon which they depend, with a threshold that depends on the instant considered. The cooling part of the dynamic model is indeed modelled as such : it does not appear to be exactly piecewise linear on Figure 4.17, because of the way the cooling degrees Δ_n^{cool} are actually computed from T_n^{cool}, see (see Bruhns et al., 2005, for the complete explanation). The heating part however is not modelled as such since the heating gradient is chosen non constant in the model (4.23). It is thus a bit of a surprise to find this familiar piecewise linear shape for the heating part, even though it is quite common for non dynamic models (see Bruhns et al., 2005, for example).

The behaviour of the dynamic model when confronted with data that include a sudden cold wave is also visible on the left of Figure 4.18 : three successive such cold waves happened in early 2010, to which the model reacted by increasing the heating part accordingly. In fact the response of the main model (4.23) is easier to observe on Figure 4.19 where only the filtered gradient of the model is displayed. The mean drift in summer should obviously be ignored: it is a direct consequence of the non symmetric random-walk prior put upon the heating gradient which forces an increase in variance and a change in mean. Note that this model artifact may easily be avoided with an autonomous model specification that does not impose sign constraints on the seasonal, heating and cooling parts : though not displayed here, the results for such a specification were studied, and it indeed led to a flat heating gradient over summer without impacting the overall performance of the model. As can be seen on the right part of Figure 4.19, the filtered standard deviation of the heating gradient gets larger during summer and smaller during winter which is logical considering that this gradient cannot be observed during summer (the temperature being above the corresponding heating threshold, it gets multiplied by zero in the end).

As a matter of fact, the behaviour of the heating gradient over summer is of little practical importance: while it is true that it cannot be observed accurately for a period of time, it also has no direct impact on the quality of the model since this is precisely the period over which the heating part of the model vanishes. The winter and mid-season time windows are of course of bigger interest : the left part of Figure 4.19 hints at the periodic nature of the heating gradient of the dynamic model with a strong gradient in winter and a weaker one in mid-season (recall that the heating gradient is negative by definition of the model, thus stronger actually means lower on Figure 4.19).

Filtered seasonal part

Even though we do not display it here, let us mention that the filtered seasonal part x_n^{season} of the model (4.23) exhibits a 1-year period with weekly cycles. Around the main periodic pattern, variations occur : more so over the winter period, for which the seasonal part is obviously not so well defined, than over the summer period. Indeed, during summer the seasonal part is the only active dynamic part of the model, while during winter the heating part also plays an important role : the estimated values of both parts over winter are thus to be interpreted with caution. Still, the filtered seasonal part seems to react correctly to the summer and winter holiday breaks (as we will outline in the next Section), although no particular information was used to flag these time windows for the model.

Because EDF customers now represents a fraction only of the French customers population (instead of the whole), the perimeter of the data varies over time due to customers departures or arrivals (but taking into account that EDF and France perimeters were actually identical until a few years ago, departures are a bit more likely). As a matter of fact, the filtered seasonal part also shows successive yearly drops from 2008 and onwards, which correspond to the financial crisis that arose in late 2008 (and that impacted the French electricity load), or planned customers' departures.

Summer break

Since holiday breaks are among the most toughest times of the year for predictions, we investigate the behaviour of the dynamic models over the summer break to show how the models cope with the difficulty.

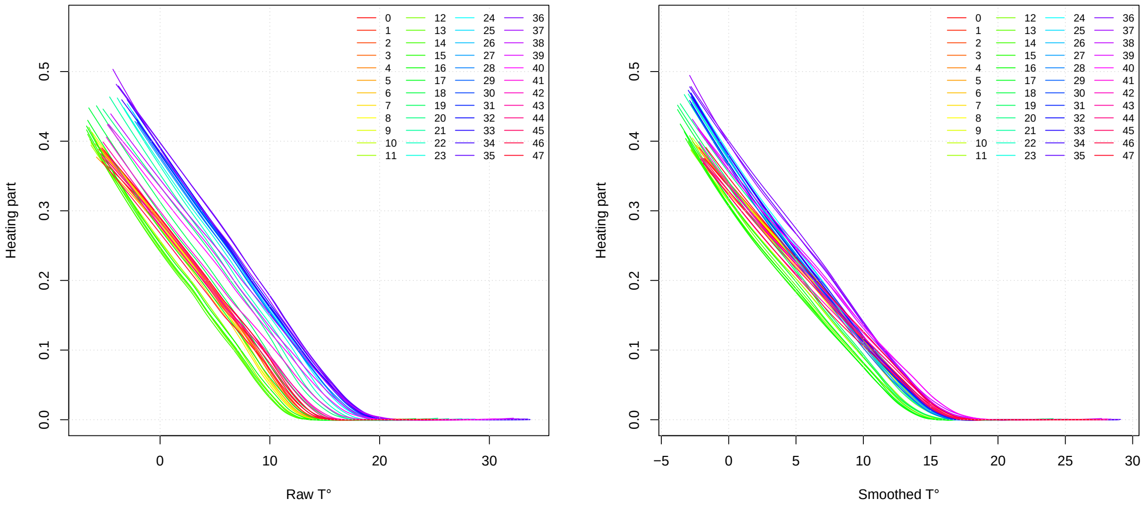


Figure 4.16: Estimated filtered heating part of the main model (4.23) against the raw temperature T_n^{raw} (left) and the smoothed temperature T_n^{heat} given to the model (right). 48 distinct colours are used, one for each of the 48 half-hours. The estimation was done via the loess function in R considering the filtered mean of the heating part against both temperatures. Only the relative heating part is shown here, i.e. the heating part divided by the maximum load observed over the whole period.

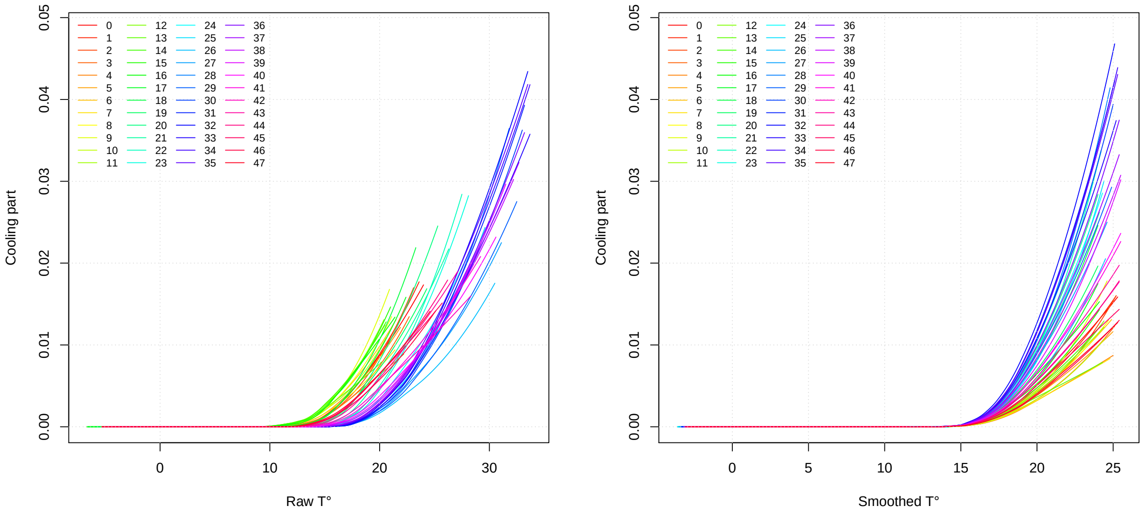


Figure 4.17: Estimated filtered cooling part of the main model (4.23) against the raw temperature T_n^{raw} (left) and the smoothed temperature T_n^{cool} (right). 48 distinct colours are used, one for each of the 48 half-hours. The estimation was done via the loess function in R considering the filtered mean of the cooling part against both temperatures. Only the relative cooling part is shown here, i.e. the cooling part divided by the maximum load observed over the whole period.

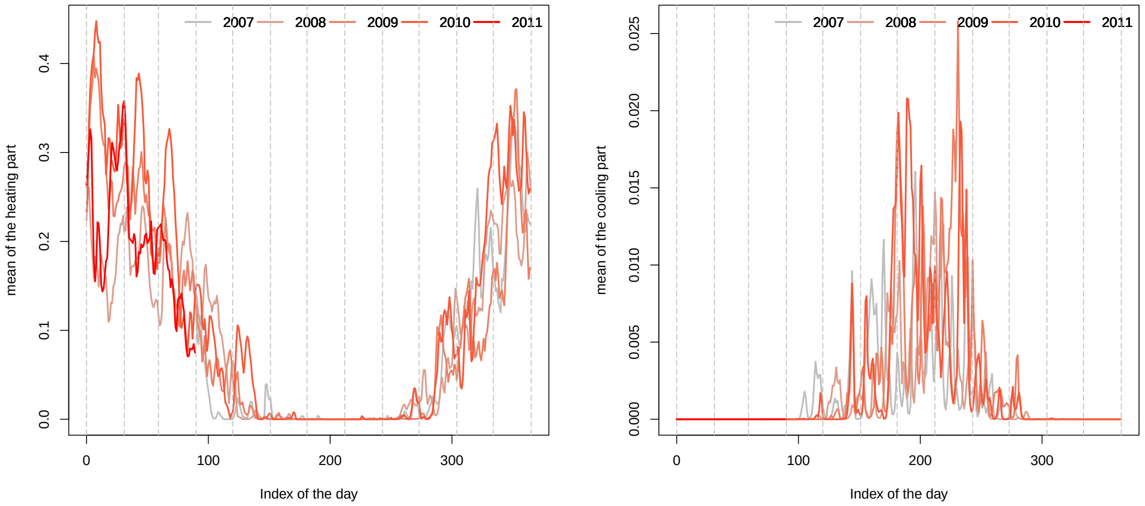


Figure 4.18: Mean of the filtered heating part x_n^{heat} (left) and cooling part x_n^{cool} (right) of the main model (4.23), averaged over 48 half-hours, as functions of the day in the calendar. The saturation of the colour used increases with each year. Only the relative heating and cooling parts are shown here, i.e. the parts divided by the maximum load observed over the whole period.

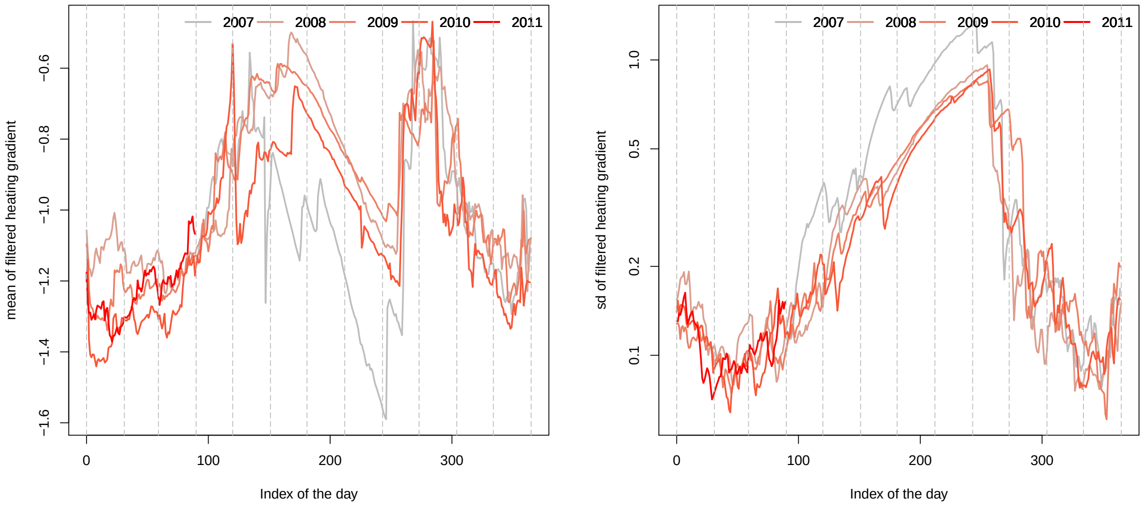


Figure 4.19: Mean (left) and standard deviation (right) of the filtered heating gradient g_n^{heat} of the main model (4.23), averaged over 48 half-hours, as a function of the day in the calendar. The saturation of the colour used increases with each year. The ordinate axis of the right figure is in log-scale. Only the relative mean and standard deviation are shown here, i.e. the mean and the standard deviation of g_n^{heat} divided by the mean of g_n^{heat} over the whole period.

Evolution of the dynamics. The Figure 4.20 shows the filtered mean of both s_n and $\sigma_{s,n}$, that rules the dynamic of s_n within the main model (4.23). As can be seen on the Figure 4.20, the model is able to filter out the summer break effectively : to allow for the sharp drop of s_n during August, the standard deviation of its dynamic $\sigma_{s,n}$ suddenly grows (becoming twice as large as usual), reflecting the brusque increase of variability of the signal over a short period of time. The model also deals with the winter break in a similar manner.

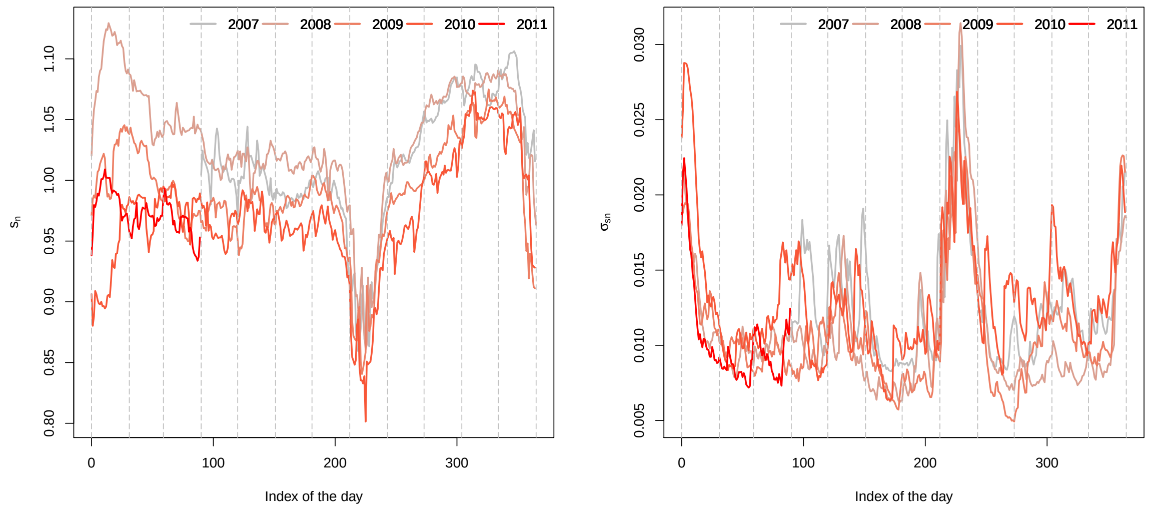


Figure 4.20: Mean of the filtered coefficients s_n (left) and $\sigma_{s,n}$ (right) of the main model (4.23) averaged over 48 half-hours, as functions of the day in the calendar. The saturation of the colour used increases with each year. Only the relative filtered means of s_n and $\sigma_{s,n}$ are shown here, i.e. the means divided by the mean of s_n over the whole period.

We have already discussed the behaviour of the heating gradient over summer, shown in Figure 4.19. Though the scaling is different, the overall behaviour of $\sigma_{g,n}$ (the coefficient that rules the dynamic of g_n^{heat}) is identical to the behaviour of the posterior standard deviation of g_n^{heat} which is why we have not represented it here. During summer the model logically loses track of anything related to the heating part, which leads to artificially increased values of $\sigma_{g,n}$.

The reasons behind the increased values of $\sigma_{s,n}$ and $\sigma_{g,n}$ during the summer break are hence entirely different. Whereas $\sigma_{s,n}$ grows to allow the model to fit data that do not match the current state, the growth of $\sigma_{g,n}$ merely reflects the lack of cold temperatures that would help estimate any of the coefficients related to the heating part of the model (4.23).

Predictive errors. Though no information is provided about the summer break (a succession of breaks mostly occurring on Mondays), we already saw that the dynamic models are able to estimate the electricity load rather correctly given the peculiar circumstances. We show in Figure 4.21 how the lack of information typically influences the summer break predictions. For a prediction horizon $\tau = 1$, the main model (4.23) provides poor predictions mostly on Mondays, weekends and August 15th (a bank holiday) but adjusts itself for the other days. Focusing on Mondays, it is easy to see that the model overpredicts the load on Mondays (although the vast majority of the corresponding instants are not detected as outliers) but recovers on the following days, i.e. on Tuesdays, as can be observed in the bottom-left corner of Figure 4.21. Similarly, for a prediction horizons $\tau = 3$ and $\tau = 5$, the model only recovers on the third or fifth day after each of the successive breaks, i.e on Thursdays or Saturdays, as is shown in the top row of Figure 4.21.

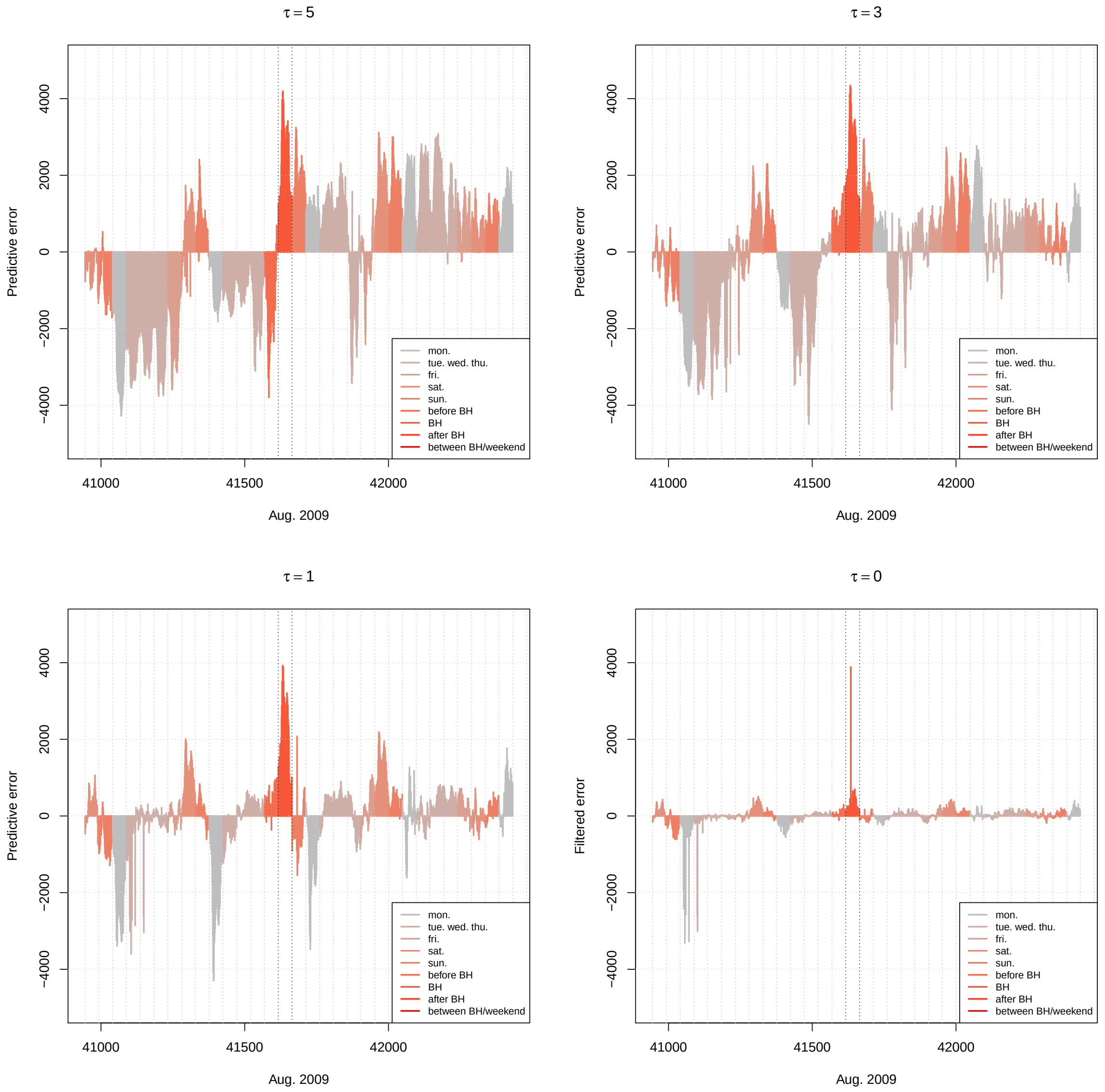


Figure 4.21: Predictive error of the main model (4.23) over the whole month of August 2009 for various horizons : $\tau = 5$ (top-left), $\tau = 3$ (top-right) $\tau = 1$ (bottom-left). The filtered error of the main model (4.23) over the same period is also shown, labeled as $\tau = 0$ (bottom-right). The colour used for each instant indicates the daytype. Vertical dotted lines are used to separate each day, with 08/15 highlighted between two darker lines.

Interestingly, we also observe that the filtered error (the difference between the actual observation and the filtered state), labelled as $\tau = 0$ in the bottom-right corner of Figure 4.21, happens to be very small, except for a few instants that were detected as outliers : 5 instants were flagged as outliers on the first Monday of August 2009, meaning that the model did not learn from the corresponding observations at these instants (recall that outliers are treated as missing in Algorithm 4.10) and thus did not re-adjust itself. For these 5 instants on the following day, the prediction was again poor which explains why outliers were still detected on 2 of these 5 instants.

A possible way to improve the quality of the forecasts for the days where breaks occur would be to taylor the transition density of one state to the next specifically for them. This requires much expertise in practise because the way the load is affected by the summer break also depends on the calendar configuration: one could for instance introduce adequately modified specifications such as

$$s_{n*} = s_{n*-1} - \mu_{n^*} + \epsilon^s_{n^*}$$

into the model where $\mu_{n^*} \in \mathbb{R}_+$ is the drop in load expected to happen at time n^*.

Comparison with a linear Gaussian state space model

In this Section, we offer a short comparison of the results obtained by Dordonnat et al. (2008) to put our own results in perspective. A dynamic model was proposed and studied by Dordonnat et al. (2008) to model the French electricity load at the national perimeter. Note that the data we consider here are different from these that they considered (different perimeters and different time ranges) which makes numerical comparisons hard.

Their model fit in the multivariate linear Gaussian state space models framework which allowed for the use of Kalman filtering and associated techniques (see Durbin and Koopman, 2001). It is actually quite a complex and rich model, compared to our own, and includes multiple regressions, some coefficients of which are allowed to vary over time : a truncated Fourier series is used to model the seasonality of the signal as in Bruhns et al. (2005) in conjunction with a stochastic trend. Local trends are also included to model the holiday breaks, and a calendar with various specific daytypes is used. Heating and cooling parts are defined as well, using fixed threshold values (15°C and 18°C) as well as fixed smoothing parameters (fixed to $\vartheta = 0.98$), and are thus very similar to the ones we use, although the heating part relied upon the use of two heating gradients (the first corresponding to the raw temperature, the other to the difference between smoothed and raw temperatures). The model was estimated using national data from 09/01/1995 to 08/30/2004, and its predictive quality was assessed from 09/01/2003 to 08/30/2004 only.

Let us first mention that the performances reported by Dordonnat et al. (2008) for their model are in accord with ours, with a one-day-ahead predictive MAPE varying around 1.30% across the 24 hours considered, and larger errors during the weekends or holiday breaks. They also found the quality of the forecasts obtained to be degrading with the predictive horizon, just as we did, and at a similar rate. Finally, the behaviour of the heating gradient that we reported in Figure 4.19 corroborates the behaviour of the heating gradients found in Dordonnat et al. (2008) (with this difference that they used a smoothing approach for the signal extraction, whereas we used a filtering approach).

Still, the dynamic models (4.23) and (4.24) that we propose are much simpler, most notably where the seasonality part is concerned: our models only include 9 daytypes and at most 2 temperatures, i.e. 10 random effects whereas the model described in Dordonnat et al. (2008) made use of more than

30 random effects. Arguably, the estimation time of our models via a particle filter takes more time than running a Kalman filter, but particle filters naturally allows for more flexibility in the definition of the model (including non-linear non-Gaussian model). Most importantly, the Algorithm 4.10 that we implemented for the estimation of our models automatically treats special instants as missing data when Dordonnat et al. (2008) explicitly and manually had to declare which data had to be considered as missing data, so as not to throw the model off. Also note that even though the model studied in Dordonnat et al. (2008) was more complex, the predictive MAPE they obtained for non regular daytypes exceeded 5% at 09:00AM and 12:00PM the two instants they focused on while dynamic models (4.23) and (4.24) had an averaged predictive MAPE of 3.34% and 3.46% for non regular daytypes (but once again keep in mind that the datasets used for their experiments and ours were different which may possibly explain part of the observed difference).

5 Conclusion

Dans ce dernier chapitre, nous proposons une synthèse de nos contributions pour chacun des sujets abordés et discutons quelques pistes possibles pour des travaux de recherche ultérieurs.

5.1 Consistance de la loi a posteriori et de l'estimateur du maximum de vraisemblance pour la régression linéaire par morceaux

5.1.1 Synthèse

Dans le chapitre 2, nous présentons les résultats asymptotiques obtenus pour un modèle de régression linéaire par morceaux, qui sert en pratique à modéliser la partie de la consommation d'électricité liée au chauffage. Grâce à la notion de pseudo-problème introduite par Sylwester (1965), nous prouvons la consistance de l'estimateur du maximum de vraisemblance et sa normalité asymptotique (avec une vitesses de convergence "standard" $\sqrt{n}$). Nos résultats complètent ceux de Feder (1975). Nous prouvons un théorème de Bernstein-von Mises pour le modèle qui généralise les résultats de Ghosh et al. (2006) à un modèle non-régulier.

5.1.2 Perspectives

Extension des résultats au modèle complet

Le théorème de Bernstein-von Mises démontré au chapitre 2 est obtenu pour le seul modèle de part chauffage. L'extension de ce résultat au modèle décrit dans Bruhns et al. (2005), par exemple pour la validation asymptotique de l'analyse statistique présentée au chapitre 3, est envisageable dans la mesure où le reste du modèle est régulier.

Développement de Edgeworth pour la loi a posteriori

Comme dans le cadre de modèles réguliers (voir Ghosh et al., 2006), le théorème de Bernstein-von Mises que nous avons démontré au chapitre 2 nous permet d'écrire le développement de Edgeworth (voir Cramér, 1946) à l'ordre 0

$$\sup_u |F_n(u) - \Phi(u)| = o_{\mathbb{P}}(1), \tag{5.1}$$

où $F_n(u)$ désigne la fonction de répartition de $\sqrt{n} J_n^{1/2} \cdot (\theta - \widehat{\theta}_n)$, en notant $J_n = \frac{1}{n} I_n$ l'information de Fisher de l'échantillon, et où Φ dénote la fonction de répartition de la loi gaussienne univariée centrée réduite.

Sous des hypothèses de régularité supplémentaires Johnson (1970) généralise ce résultat pour des modèles réguliers en prouvant le développement presque sûr suivant

$$\sup_u \left| F_n(u) - \Phi(u) - \phi(u) \sum_{j=1}^{k} \psi_j(u; X_1, \ldots, X_n) n^{-j/2} \right| = \mathrm{O}\left(n^{-(k+1)/2}\right), \tag{5.2}$$

où ϕ dénote la densité de la loi gaussienne univariée centrée réduite et où chacunes des fonctions $\psi_j(u; X_1, \ldots, X_n)$ est un polynôme en u dont les coefficients sont bornés en $X_1, \ldots, X_n$.

Le concept de pseudo-problème introduit par Sylwester (1965) pourrait être réutilisé pour contourner la difficulté de la non-différentiabilité du modèle (2.1) et obtenir dans un premier temps un développement de Edgeworth pour la loi a posteriori du pseudo-problème. La généralisation au problème complet pourrait alors s'effectuer dans un second temps.

Lois a priori "matching"

Nous reprenons ci-dessous les notations introduites en section 1.2 pour les lois "matching". Datta and Mukerjee (2004) montrent sous des hypothèses générales de régularité le développement suivant

$$\mathbb{P}_\theta\big(\theta_1 \leqslant q_\alpha^\pi(x)\big) = \alpha + n^{-\frac{1}{2}} \frac{\phi(z)}{\pi(\theta)} \Delta_1(\pi, \theta) + \mathrm{O}\left(n^{-1}\right), \tag{5.3}$$

où z désigne le quantile d'ordre α et ϕ la densité de la loi gaussienne univariée centrée réduite et où

$$\Delta_1(\pi, \theta) = \sum_{j=1}^{d} \frac{\partial}{\partial \theta_j} \left\{ \pi(\theta) I^{j1}(\theta) \big(I^{11}(\theta)\big)^{-\frac{1}{2}} \right\}. \tag{5.4}$$

Pour le modèle de régression linéaire par morceaux, le calcul d'un tel développement se heurte à nouveau au problème de la non différentiabilité de la vraisemblance. La matrice d'information de Fisher qui apparaît dans l'expression (5.4) et sur laquelle se base la caractérisation $\Delta_1(\pi, \theta) = 0$ des lois "matching" démontrée par Datta and Mukerjee (2004), n'est pas définie en tout point de l'espace des paramètres et varie avec le nombre d'observations et la variable exogène. L'obtention d'un développement semblable (5.3) n'est donc pas immédiate.

A nouveau, le concept de pseudo-problème peut vraisemblablement être réutilisé pour contourner la difficulté de la non-différentiabilité du modèle. Un premier objectif théorique consisterait donc à montrer l'équivalence entre les notions de lois "matching" pour le pseudo-problème et de lois "matching" pour le problème complet.

L'extension des résultats de Datta and Mukerjee (2004) au pseudo-problème constitue un deuxième objectif théorique. Dans le cadre du modèle (2.1), les observations ne sont pas i.i.d. et l'information de Fisher empirique n'est plus constante. Nous pensons donc qu'il est nécessaire de considérer l'information de Fisher asymptotique (définie en (2.7), sous l'hypothèse A1) comme le font par exemple Philippe and Rousseau (2002) dans le cadre de la longue mémoire. Ceci pourrait permettre de lever la difficulté liée à la présence d'une variables exogène, sous réserve de connaître l'expression de F la limite des fonctions de répartition empirique F_n de la variable exogène qui intervient dans l'hypothèse A1 donnée en page 28.

Dans le cadre de la prévision de consommation d'électricité, il nous paraît également pertinent de considérer des lois "matching" vis-à-vis de la loi prédictive. Il s'agit alors de caractériser les lois a priori pour lesquelles le niveau fréquentiste des intervalles de crédibilité de la loi prédicitive (et non plus de la loi a posteriori) tend plus rapidement vers le taux de couverture bayésien. Sweeting (2008) discute par exemple l'existence et l'unicité de telles lois dans le cadre de modèles multivariés réguliers.

5.2 PRÉVISION EN SITUATION D'HISTORIQUE COURT

5.2.1 Synthèse

Dans le chapitre 3 nous développons une méthode de construction de loi a priori pour un modèle en situation d'historique court afin d'améliorer la qualité des prévisions. A partir d'un jeu de données long $\mathcal{A}$ et supposé semblable au jeu de données court $\mathcal{B}$, nous construisons une loi a priori hiérarchique en introduisant des hyperparamètres dont le rôle est de modéliser la ressemblance entre les deux jeux de données $\mathcal{A}$ et $\mathcal{B}$. Nous montrons à travers des applications, sur des données simulées et réelles, l'apport d'une telle loi a priori sur la qualité des prévisions du modèle utilisé.

5.2.2 Perspectives

Estimations simultanées

La méthode que nous proposons dans le chapitre 3 repose essentiellement sur deux estimations successives du modèle : estimation sur le jeu de données $\mathcal{A}$ à l'aide d'une loi non informative, puis estimation sur le jeu de données $\mathcal{B}$ en incorporant une partie de l'information apprise sur $\mathcal{A}$. Ceci permet notamment de réutiliser l'estimation effectuée sur le jeu de données $\mathcal{A}$ pour plusieurs jeux de données $\mathcal{B}_j$ $(j > 1)$.

Un prolongement possible des travaux présentés dans le chapitre 3 consiste à effectuer une estimation simultanée du modèle sur tous les jeux de donneés $\mathcal{B}_j$ (et non plus des estimations séparées) afin de mutualiser l'information contenue dans chacun d'eux, sous réserve qu'ils partagent tous approximativement la même ressemblance avec le jeu de données $\mathcal{A}$. Il devient alors nécessaire de modéliser la relation entre les jeux de données $\mathcal{B}_j$, ce qui constitue une difficulté de cette approche.

Un autre prolongement envisageable consiste à considérer une estimation conjointe du modèle sur les deux jeux de données $\mathcal{A}$ et $\mathcal{B}$ (au lieu d'estimer séparément le modèle sur $\mathcal{A}$ puis sur $\mathcal{B}$) : dans ce cas, il est nécessaire de modéliser la relation entre $\mathcal{A}$ et $\mathcal{B}$. Signalons néanmoins qu'une telle démarche introduit implicitement un effet de $\mathcal{B}$ sur $\mathcal{A}$, qui va à l'encontre de l'approche originale : dans le chapitre 3, seul $\mathcal{B}$ est considéré comme un jeu de données d'intérêt et la démarche proposée s'articule par conséquent autour d'un effet unilatéral de $\mathcal{A}$ sur $\mathcal{B}$.

Classification des facteurs de ressemblance

Puisque le comportement des différents facteurs de ressemblance k_j du modèle n'est pas homogène, nous envisageons une séparation des coefficients k_j en plusieurs classes.

Nous pensons en premier lieu à une séparation des facteurs de ressemblance k en deux blocs $k = (k_1, k_2)$, où le premier bloc k_1 regroupe les facteurs de ressemblance des paramètres influencés par le volume de la population (coefficients de Fourier, gradients), et où le second les coefficients de ressemblance des paramètres non influencés par le volume de la population (forme de jours, seuil de chauffage). Pour une telle classification nous suggérons, par exemple, d'utiliser une loi a priori sur k_2 (possiblement hiérarchique) centrée en 1 et une loi a priori sur k_1 centrée en ρ où ρ désigne le rapport moyen entre les signaux des deux jeux de données $\mathcal{A}$ et $\mathcal{B}$.

L'approche que nous venons de décrire se base sur des classes de facteurs de ressemblance déjà connues mais nous pensons qu'une approche de classification supervisée ou non supervisée (à partir de lois a priori formulée comme des mélanges) pourrait également s'avérer intéressante.

Compréhension fine de l'apport d'information

Les différents résultats du chapitre 3 montrent l'apport indéniable de l'information a priori dans les processus d'estimation puis de prévision du modèle, mais il est difficile d'en préciser exactement l'origine. La formulation de la loi a priori pour le jeu de données $\mathcal{B}$ sous la forme d'une loi gaussienne hiérarchique

$$\mathcal{N}(K\mu^{\mathcal{A}}, l^{-1}\Sigma^{\mathcal{A}})$$

permet au modèle d'adapter la moyenne $\mu^{\mathcal{A}}$ pour se caler autour des paramètres du jeu de données $\mathcal{B}$. Nous pensons qu'une grande partie de l'apport d'information a priori s'effectue à travers la forme de la matrice de covariance choisie $l^{-1}\Sigma^{\mathcal{A}}$. Elle impose a priori sur le jeu de données $\mathcal{B}$ la même structure de corrélation que celle obtenue a posteriori sur $\mathcal{A}$. Afin de vérifier cette conjecture, nous suggérons d'utiliser une décomposition matricielle de Σ appartenant à $S_d^+(\mathbb{R})$ l'ensemble des matrices symétriques définies positives de dimension d, en trois composantes : volume v, forme S et direction D (la terminologie adoptée faisant référence à la géométrie des ellipsoïdes de confiance associés à une loi gaussienne de matrice de covariance Σ)

$$\Sigma = v \cdot DSD' \tag{5.5}$$

où $v = \{\det \Sigma\}^{1/d}$, S est une matrice diagonale dont les d coefficients diagonaux $S_{11} \geqslant \ldots \geqslant S_{dd} > 0$ vérifient $\prod_{i=1}^{d} S_{ii} = 1$ et D est une matrice groupe orthogonal $\mathrm{O}_d(\mathbb{R})$.

Supposées totalement inconnues, les trois composantes v, S et D de la décomposition (5.5) correspondent chacunes à 1, $d-1$ et $\frac{1}{2}d(d-1)$ degrés de liberté, et permettent d'identifier Σ de manière presque unique. Cette décomposition permet de choisir une loi a priori pour le jeu de données $\mathcal{B}$ de la forme

$$\mathcal{N}(K\mu^{\mathcal{A}}, v \cdot DSD')$$

en autorisant les paramètres v, D et S à varier autour des valeurs $v^{\mathcal{A}}$, $D^{\mathcal{A}}$ et $S^{\mathcal{A}}$ (associée à la décomposition (5.5) de la matrice $\Sigma^{\mathcal{A}}$).

Dans le chapitre 3, la méthode que nous proposons est basée sur cette décomposition avec $D = D^{\mathcal{A}}$ et $S = S^{\mathcal{A}}$: l'hyperparamètre l autorise un degré de liberté sur le volume v.

L'introduction de degrés de liberté supplémentaires dans le modèle peut permettre de préciser l'importance des différentes composantes, et d'autoriser un contrôle plus fin de la contribution a priori provenant du jeu de données $\mathcal{A}$. Une approche partiellement informative est par exemple envisageable, en fixant

$$D = R \cdot D^{\mathcal{A}},$$

où R est une matrice du groupe spécial orthogonal $\mathrm{SO}_d(\mathbb{R})$ a priori "proche" de la matrice identité (par exemple une matrice de rotation d'angle faible, ou un produit de telles matrices).

5.3 APPLICATION DU FILTRAGE PARTICULAIRE À LA PRÉVISION DE CONSOMMATION D'ÉLECTRICITÉ

5.3.1 Synthèse

Dans le chapitre 4 nous nous intéressons à des modèles à espace d'états pour obtenir des prévisions en ligne (i.e. mises à jour au cours du temps, au fur et à mesure de l'arrivée des observations). Nous abordons les problèmes inhérents à l'utilisation de filtres particulaires, discutons diverses solutions développées dans la littérature pour y répondre, et fournissons les algorithmes détaillés permettant une implémentation rapide. Nous proposons également une méthode simple pour rendre les filtres particulaires plus robustes vis-à-vis de données atypiques (i.e. peu probables étant donné l'état actuel du modèle). Les deux modèles que nous choisissons d'étudier font principalement appel à deux dynamiques (saisonnalité et part chauffage) : le premier modèle repose sur l'utilisation d'une température lissée précalculée pour un modèle opérationnel tandis que le second détermine de manière autonome sa propre température de référence comme un mélange entre la température brute et une température très lissée. Nous montrons que ces deux modèles fournissent globalement des prévisions compétitives avec les prévisions opérationnelles. L'autonomie du second modèle s'obtient au prix d'une qualité de prévision légèrement dégradée par rapport au premier modèle.

5.3.2 Perspectives

Amélioration de la technique d'estimation

Comme nous l'avons signalé au cours du chapitre 4, les données atypiques constituent un point faible reconnu des filtres particulaires les plus basiques. De nombreux efforts sont investis dans la littérature pour atténuer cette difficulté, par exemple au travers d'un choix adéquat de loi instrumentale dans l'étape d'échantillonnage d'importance (voir la discussion en section 4.3.6). Rappelons que dans l'algorithme 4.10 retenu, nous choisissons d'utiliser la loi a priori comme loi instrumentale et compensons la sensibilité du filtre aux données atypiques par leur traitement en tant que données manquantes. L'implémentation d'autres choix de lois instrumentales, par exemple basées sur des filtres de Kalman (voir van der Merwe et al., 2001; Wan and van der Merwe, 2000), représente donc un enjeu important pour exploiter toute l'information disponible dans le contexte des modèles dynamiques, et éviter d'ignorer la moindre observation.

Modifications mineures du modèle

Deux modifications du modèle peuvent être rapidement envisagées pour améliorer la qualité de prévision sur les jours les plus difficiles.

L'utilisation d'un calendrier offrant des types de jours plus détaillés constitue une première piste à explorer et devrait par exemple permettre de corriger sensiblement les mauvaises prévisions obtenues sur les jours fériés.

Une deuxième piste possible est l'inclusion d'un lissage exponentiel de la température afin de mieux représenter l'effet de la température sur la consommation d'électricité. Le coefficient de lissage pourra alors être estimé en ligne qu'il soit choisi fixe au cours du temps ou dynamique.

Modifications majeures du modèle

Améliorer les prévisions autour des périodes de rupture telles que les congés d'été et d'hiver représente un enjeu opérationnel fort. Une piste d'exploration possible pour y parvenir consiste à introduire de nouvelles variables exogènes (appelées interventions) dans la définition du modèle dynamique pour réajuster le signal prévu au niveau plus faible attendu au début des périodes de rupture.

Le traitement simultané des 48 instants de la journée (à partir d'un modèle vectoriel), avec introduction de dépendances entre les instants, est aussi envisageable, comme l'a déjà montré Dordonnat (2009) avec une modélisation dynamique du signal journalier à partir de fonctions splines.

Une méthode alternative pour traiter les différents instants de la journée consiste à considérer le signal comme une série temporelle univariée et à la modéliser comme une série périodiquement corrélée (voir Hurd and Miamee, 2007; Serpedin et al., 2005, par exemple).

La modélisation dynamique proposée au chapitre 4 permet d'estimer des intervalles de confiance sur les prévisions et la longueur de ces intervalles de confiance varie de manière naturelle au cours du temps. En considérant l'erreur de prévision opérationnelle actuelle comme nouveau signal d'étude, l'utilisation d'un modèle dynamique pourrait permettre de proposer un correctif à la prévision existante pour améliorer sa qualité et également disposer d'intervalles de confiance.

Bibliographie

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. 74

Andrieu, C., de Freitas, N., and Doucet, A. (1999). Sequential MCMC for Bayesian Model Selection. In *IEEE Higher Order Statistics Workshop, Ceasarea*, pages 130–134. 117

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72(3):269–342. 117

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Soc. of London*, 53:370–418. 6

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B*, 41(2):113–147. 10

Billingsley, P. (1995). *Probability and Measure*. Wiley, 3rd edition. 56, 59

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, 2nd edition. 29

Breiman, L. (1992). *Probability*. SIAM. 50

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*. 4, 27, 72, 74, 89, 119, 120, 126, 137, 138, 143, 145

Bunn, D. and Farmer, E. (1985). *Comparative Models For Electrical Load Forecasting*. John Wiley, New York. 71

Cappé, O., Moulines, E., and Ryden, T. (2010). *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer. 103

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *Radar, Sonar and Navigation, IEE Proceedings*, 146(1):2–7. 112

Chen, Z. (2003). Bayesian Filtering : From Kalman Filters to Particles Filters, and Beyond. 103, 110, 111, 112, 116

Chopin, N. (2004). Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *The Annals of Statistics*, 32(6):2385–2411. 117

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2012). SMC^2: an efficient algorithm for sequential analysis of state-space models. *Journal of the Royal Statistical Society: Series B*. 102, 117

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *Information Theory, IEEE Transactions on*, 36(3):453–471. 10

Congdon, P. D. (2003). *Applied Bayesian Modelling*. Wiley Series in Probability and Statistics. John Wiley & Sons. 9

Congdon, P. D. (2010). *Bayesian Random Effect and Other Hierarchical Models: An Applied Perspective*. Chapman and Hall/CRC. 9

Cornebise, J. (2009). *Méthodes de Monte Carlo Séquentielles Adaptatives*. PhD thesis, Université Pierre et Marie Curie. 110, 112, 127

Cottet, R. and Smith, M. (2003). Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849. 4, 72, 74

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton mathematical series. Princeton University Press. 145

Crisan, D. and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 50(3):736–746. 117

Cugliari, J. (2011). *Prévision non paramétrique de processus à valeurs fonctionnelles, application à la consommation d'électricité*. PhD thesis, Univérsité Paris Sud XI. 2, 4, 71

Dacunha-Castelle, D. (1978). Vitesse de convergence pour certains problèmes statistiques. In *École d'Été de Probabilités de Saint-Flour, VII (Saint-Flour, 1977)*, volume 678 of *Lecture Notes in Math.*, pages 1–172. Springer, Berlin. 19, 37, 38

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer, 1st edition. 6, 114

Datta, G. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Springer. 8, 10, 146

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer, 1 edition. 12

Dordonnat, V. (2009). *State-space modelling for high frequency data*. PhD thesis, Vrije Universiteit Amsterdam. 2, 4, 72, 102, 120, 150

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. 24, 26, 72, 103, 143, 144

Douc, R. and Cappe, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 64–69. IEEE. 112

Douc, R. and Moulines, E. (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist.*, 36(5):2344–2376. 117

Douc, R. and Moulines, E. (2012). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Preprint*. arXiv:1203.6898. 117

Doucet, A. (1998). On sequential Simulation-Based methods for bayesian filtering. Technical Report CUED/F-INFENG/TR. 310, Cambridge University Department of Engineering. 109, 111, 114

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice (Statistics for Engineering and Information Science)*. Springer, 1st edition. 102, 113

Doucet, A., Godsill, S., and Andrieu, C. (2000). On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statistics and Computing*, 10:197–208. 109, 110, 111, 112, 116

Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: fifteen years later. 113

Durbin, S. and Koopman, S. (2001). *Series Analysis by State Space Methods*. Oxford Statistical Science Series. 102, 109, 143

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. 4, 72

Fan, J. and Yao, Q. (2005). *Non linear Time Series: Nonparametric and Parametric Methods*. Springer. 87

Feder, P. I. (1975). On asymptotic distribution theory in segmented regression problems – identified case. *The Annals of Statistics*, 3(1):49–83. 17, 18, 27, 28, 31, 36, 47, 145

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. 72

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472. 125

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741. 14

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica*, 57(6):1317–1339. 16, 106

Ghosal, S. and Samanta, T. (1995). Asymptotic behaviour of Bayes estimates and posterior distributions in multiparameter nonregular cases. *Math. Methods Statist.*, 4(4):361–388. 37

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis, Theory and Methods*. Springer. 17, 18, 27, 28, 31, 33, 34, 37, 63, 145

Ghosh, J. K., Ghosal, S., and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical decision theory and related topics, V (West Lafayette, IN, 1992)*, pages 183–199. Springer, New York. 37

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer. 6, 30

Gilks, W. R. and Berzuini, C. (2001). Following a moving target – Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B*, pages 127–146. 102, 113

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113. 111, 113

Goude, Y. (2008). *Mélange de prédicteurs, Application à la prévision de consommation d'électricité*. PhD thesis, Univérsité Paris Sud XI. 2, 4, 72

Gustafsson, F., Gunnarsson, F., Bergman, N., Forssell, U., Jansson, J., Karlsson, R., and Nordlund, P. J. (2002). Particle filters for positioning, navigation, and tracking. *IEEE Transactions on Signal Processing*, 50(2):425–437. 102

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. 2, 71

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109. 13

Higuchi, T. (2001). Self-organizing time series model. In *Sequential Monte Carlo Methods in Practice*, pages 429–444. Springer-Verlag. 117

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. 72

Hoare, C. A. R. (1962). Quicksort. *The Computer Journal*, 5(1):10–16. 111, 127

Hu, X.-L., Schön, T. B., and Ljung, L. (2008). A basic convergence result for particle filtering. *IEEE Transactions on Signal Processing*, 56(4):1337–1348. 115, 117

Hu, X.-L., Schön, T. B., and Ljung, L. (2011). A general convergence result for particle filtering. *IEEE Transactions on Signal Processing*, 59(7):3424–3429. 115

Hurd, H. and Miamee, A. (2007). *Periodically Correlated Random Sequences: Spectral Theory and Practice*. Wiley Series in Probability and Statistics. John Wiley & Sons. 150

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 88(3). 87

Ibragimov, I. A. and Khasminskii, R. Z. (1981). *Statistical estimation : Asymptotic theory*. Springer-Verlag, New York. 28, 37

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461. 10

Johansen, A., Doucet, A., and Davy, M. (2008). Particle methods for maximum likelihood estimation in latent variable models. *Statistics and Computing*, 18(1):47–57. 102

Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.*, 41:851–864. 146

Kantas, N., Doucet, A., Singh, S., and Maciejowski, J. M. (2009). Overview of Sequential Monte Carlo methods for parameter estimation on general state space models. In *15th IFAC Symposium on System Identification (SYSID), Saint-Malo, France.* (invited paper). 102, 117

Karlsson, R. (2005). *Particle Filtering for Positioning and Tracking Applications*. Linköping Studies in Science and Technology. Thesis No 924, Linköping Universitet. 102

Kass, R. E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, 91(435):1343–1370. 9

Kitagawa, G. (1996). Monte carlo filter and smoother for Non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25. 112, 117

Kitagawa, G. (1998). A Self-Organizing State-Space Model. *Journal of the American Statistical Association*, 93(443):1203–1215. 117

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288. 110

Laplace, P.-S. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie Royale des Sciences présentés par divers savans*, 6:621–656. 9

Launay, T., Philippe, A., and Lamarche, S. (2012a). Consistency of the posterior distribution and MLE for piecewise linear regression. *Electron. J. Statist.*, 6:1307–1357. 17, 79, 122

Launay, T., Philippe, A., and Lamarche, S. (2012b). Construction of an informative hierarchical prior distribution. Application to electricity load forecasting. *Preprint*. arXiv:1109.4533. 27, 120, 121

Lee, J.-Y., Payandeh, S., and Trajković, L. (2010). The internet-based teleoperation: Motion and force predictions using the particle filter method. *ASME Conference Proceedings*, 2010(44458):765–771. 102

Lehmann, E. L. (2004). *Elements of large-sample theory*. Springer Texts in Statistics. Springer-Verlag, New York. 37

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, 2nd edition. 10

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In Freitas, D. and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York. 102, 117

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, corrected edition. 102, 110

Liu, J. S. and Chen, R. (1995). Blind Deconvolution via Sequential Imputations. *Journal of the American Statistical Association*, 90(430). 110

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association*, 93:1032–1044. 112

Loève, M. (1991). *Proability Theory I*. Springer, 4th edition. 55

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337. 124, 125

Marin, J.-M. and Robert, C. (2007). *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. Springer. 12, 91, 103

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. 4, 72

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092. 13

Moral, P. D. and Guionnet, A. (1999). Central Limit Theorem for Non Linear Filtering and Interacting Particle Systems. *Ann. Appl. Probab*, 9:275–297. 117

Oudjane, N. (2000). *Stabilité et approximations particulaires en filtrage non linéaire, Application au pistage*. PhD thesis, Université de Rennes 1. 114

Oudjane, N. and Rubenthaler, S. (2005). Stability and uniform particle approximation of nonlinear filters in case of non ergodic signals. *Stochastic Analysis and applications*, 23:421–448. 117

Philippe, A. and Rousseau, J. (2002). Non-informative priors in the case of gaussian long-memory processes. *Bernoulli*, 8(4):451–473. 10, 146

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. In *Proceedings of the Sixteenth International Conference on Intelligent System Application to Power Systems (ISAP)*. 4

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599. 116

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. 124, 125

Polya, G. and Szegö, G. (2004). *Problems and Theorems in Analysis I*. Springer. 29

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. 4, 71

Robert, C. P. (1996). *Methodes de Monte Carlo par chaines de Markov*. Economica. 103

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Verlag, New York, 2nd edition. 6, 7, 8

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd edition. 13, 15, 103

Robert, C. P. and Casella, G. (2009). *Introducing Monte Carlo Methods with R*. Springer Verlag, 1st edition. 13, 16

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367. 94, 97

Rossi, V. (2004). *Filtrage non linéaire par noyaux de convolution, Application à un procédé de dépollution biologique*. PhD thesis, Ecole Nationale Supérieure Agronomique de Montpellier. 102, 114, 117

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In Bernardo, M. H., Degroot, K. M., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 3*. Oxford University Press. 16

Rui, Y. and Chen, Y. (2001). Better proposal distributions: Object tracking using unscented particle filter. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA*, pages 786–793. IEEE Computer Society. 102

Sareen, S. (2003). Reference bayesian inference in non-regular models. *Journal of Econometrics*, 113:265–288. 28

Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear Regression (Wiley Series in Probability and Statistics)*. Wiley-Interscience. 76

Serpedin, E., Panduru, F., Sari, I., and Giannakis, G. B. (2005). Bibliography on cyclostationarity. *Signal Processing*, 85(12):2233–2303. 150

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1st edition. 113, 114

Smith, M. (2000). Modeling and short-term forecasting of new south wales electricity system load. *Journal of Business & Economic Statistics*, 18:465–478. 2, 71

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97:1141–1153. 4, 72

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. 4, 71

Sweeting, T. J. (2008). On predictive probability matching priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 46–59. Inst. Math. Statist., Beachwood, OH. 147

Sylwester, D. L. (1965). On maximum likelihood estimation for two-phase linear regression. Technical Report 11, Department of Statistics, Stanford University. 18, 28, 31, 37, 145, 146

Taylor, J., Demenezes, 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):1–16. 2, 71

Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8):799–805. 2, 71, 87

Taylor, J. W. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19(1):57–70. 4, 72

Taylor, J. W. and McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on european data. *Power Systems, IEEE Transactions on*, 22(4):2213–2219. 2, 71

van der Merwe, R., de Freitas, N., Doucet, A., and Wan, E. (2001). The Unscented Particle Filter. In *Advances in Neural Information Processing Systems 13*. 116, 149

van der Vaart, A. W. (2000). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press. 11, 13, 16

Vavoulis, D. V., Straub, V. A., Aston, J. A. D., and Feng, J. (2012). A Self-Organizing State-Space-Model Approach for Parameter Estimation in Hodgkin-Huxley-Type Models of Single Neurons. *PLoS Comput Biol*, 8(3):e1002401. 102

Wan, E. A. and van der Merwe, R. (2000). The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 153–158. IEEE. 102, 116, 149

Résumé

Dans ce manuscrit, nous développons des outils de statistique bayésienne pour la prévision de consommation d'électricité en France. Nous prouvons tout d'abord la normalité asymptotique de la loi a posteriori (théorème de Bernstein-von Mises) pour le modèle linéaire par morceaux de part chauffage et la consistance de l'estimateur de Bayes. Nous décrivons ensuite la construction d'une loi a priori informative afin d'améliorer la qualité des prévisions d'un modèle de grande dimension en situation d'historique court. A partir de deux exemples impliquant les clients non télérelevés de EDF, nous montrons notamment que la méthode proposée permet de rendre l'évaluation du modèle plus robuste vis-à-vis du manque de données. Nous proposons enfin un nouveau modèle dynamique, non-linéaire, pour prévoir la consommation d'électricité en ligne. Nous construisons un algorithme de filtrage particulaire afin d'estimer ce modèle et comparons les prévisions obtenues aux prévisions opérationnelles utilisées au sein d'EDF.

Mots-clés

Consommation d'électricité ; filtrage particulaire ; historique court ; loi a priori hiérarchique ; modèle dynamique ; prévision ; régression linéaire par morceaux ; théorème de Bernstein-von Mises

Abstract

In this manuscript, we develop Bayesian statistics tools to forecast the French electricity load. We first prove the asymptotic normality of the posterior distribution (Bernstein-von Mises theorem) for the piecewise linear regression model used to describe the heating effect and the consistency of the Bayes estimator. We then build a a hierarchical informative prior to help improve the quality of the predictions for a high dimension model with a short dataset. We typically show, with two examples involving the non metered EDF customers, that the method we propose allows a more robust estimation of the model with regard to the lack of data. Finally, we study a new nonlinear dynamic model to predict the electricity load online. We develop a particle filter algorithm to estimate the model et compare the predictions obtained with operationnal predictions from EDF.

Keywords

Bernstein-von Mises theorem ; dynamic model ; electricity load ; forecasting ; hierarchical prior distribution ; particle filter ; piecewise linear regression ; short dataset