Régression linéaire bayésienne sur données fonctionnelles

Paul-Marie Grollemund

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Biostatistique

École doctorale I2S - Information, Structures, Systèmes

Unité de recherche UMR 5149 - IMAG - Institut Montpelliérain Alexander Grothendieck

Régression linéaire bayésienne sur données fonctionnelles

Présentée par Paul-Marie GROLLEMUND

Le 22 Novembre 2017

Sous la direction de Christophe ABRAHAM
et Pierre PUDLO

Devant le jury composé de

Christophe ABRAHAM, Professeur, SupAgro-INRA	Directeur
Meïli BARAGATTI, Maître de conférences, SupAgro-INRA	Examinatrice
Pierre DRUILHET, Professeur des universités, Université Blaise Pascal	Rapporteur
André MAS, Professeur des universités, Université de Montpellier	Examinateur
Éric PARENT, Professeur, AgroParisTech	Président du jury
Anne PHILIPPE, Professeur des universités, Université de Nantes	Examinatrice
Pierre PUDLO, Professeur des universités, Aix-Marseille Université	Co-Directeur
Judith ROUSSEAU, Professeure des universités, Université Paris Dauphine	Rapporteure

UNIVERSITÉ DE MONTPELLIER

Vous me dites, Monsieur, que j'ai mauvaise mine,
Qu'avec cette vie que je mène, je me ruine,
Que l'on ne gagne rien à trop se prodiguer,
Vous me dites enfin que je suis fatigué.

Oui je suis fatigué, Monsieur, mais je m'en flatte.
J'ai tout de fatigué, la voix, le coeur, la rate,
Je m'endors épuisé, je me réveille las,
Mais grâce à Dieu, Monsieur, je ne m'en soucie pas.
Ou quand je m'en soucie, je me ridiculise.
La fatigue souvent n'est qu'une vantardise.
On n'est jamais aussi fatigué qu'on le croit !
Et quand cela serait, n'en a-t-on pas le droit ?

Je ne vous parle pas des tristes lassitudes,
Qu'on a lorsque le corps harassé d'habitude,
N'a plus pour se mouvoir que de pâles raisons...
Lorsqu'on a fait de soi son unique horizon...
Lorsqu'on a rien à perdre, à vaincre, ou à défendre...
Cette fatigue-là est mauvaise à entendre ;
Elle fait le front lourd, l'oeil morne, le dos rond.
Et vous donne l'aspect d'un vivant moribond...

Mais se sentir plier sous le poids formidable
Des vies dont un beau jour on s'est fait responsable,
Savoir qu'on a des joies ou des pleurs dans ses mains,
Savoir qu'on est l'outil, qu'on est le lendemain,
Savoir qu'on est le chef, savoir qu'on est la source,
Aider une existence à continuer sa course,
Et pour cela se battre à s'en user le coeur...
Cette fatigue-là, Monsieur, c'est du bonheur.

Et sûr qu'à chaque pas, à chaque assaut qu'on livre,
On va aider un être à vivre ou à survivre ;
Et sûr qu'on est le port et la route et le gué,
Où prendrait-on le droit d'être trop fatigué ?
Ceux qui font de leur vie une belle aventure,
Marquent chaque victoire, en creux, sur la figure,
Et quand le malheur vient y mettre un creux de plus
Parmi tant d'autres creux il passe inaperçu.

La fatigue, Monsieur, c'est un prix toujours juste,
C'est le prix d'une journée d'efforts et de lutte.
C'est le prix d'un labeur, d'un mur ou d'un exploit,
Non pas le prix qu'on paie, mais celui qu'on reçoit.
C'est le prix d'un travail, d'une journée remplie,
C'est la preuve, Monsieur, qu'on vit avec la vie.

Quand je rentre la nuit et que ma maison dort,
J'écoute mes sommeils, et là, je me sens fort ;
Je me sens tout gonflé de mon humble souffrance,
Et ma fatigue alors est une récompense.

Et vous me conseillez d'aller me reposer !
Mais si j'acceptais là, ce que vous proposez,
Si je m'abandonnais à votre douce intrigue...
Mais je mourrais, Monsieur, tristement... de fatigue.

L'éloge de la fatigue de Robert Lamoureux

Remerciements

Mes premiers remerciements vont naturellement à mes directeurs de thèse. Meïli, Pierre et Christophe, bien qu'un encadrement à trois encadrants soit complexe à gérer, vous avez toujours fait preuve d'une très grande patience et d'une grande bienveillance à mon égard. Vous avez su gérer, mes périodes de flottements, mes égarements scientifiques, mon humeur, mes doutes, mes appréhensions, mes espérances, ma naïveté et bien souvent mon ignorance. Vous m'avez chacun apporté, scientifiquement et personnellement, beaucoup de choses complémentaires. Sans compter l'apport important de vos connaissances scientifiques à cette thèse, vous m'avez appris à faire de la recherche, à structurer une réflexion, à être méthodique, à être le plus clair possible et à ne jamais me reposer sur mes lauriers. Si mes études m'ont permis d'acquérir une base de connaissances, ce n'est rien face à la quantité de choses que vous m'avez inculquée. Encore une fois, merci pour ces années et votre engagement !

Je voudrais aussi exprimer ma gratitude à Judith Rousseau et Pierre Druilhet d'avoir accepté de rapporter ma thèse. C'est un honneur que mes travaux aient été relus par des chercheurs de votre envergure. Merci d'avoir pris de votre temps pour relire ce manuscrit et merci pour vos commentaires encourageants ! Je remercie aussi tous les membres de mon jury d'avoir accepté sans hésitation de faire partie du jury de cette thèse. Merci à Anne Philippe pour nos discussions très intéressantes, à André Mas pour son soutien et sa bienveillance, et à Eric Parent de me faire l'honneur d'être présent pour ma soutenance.

J'en profite également pour remercier Sophie Donnet et Alice Cleynen d'avoir participé à mes comités de suivi de thèse. Vous m'avez écouté, conseillé, soutenu. Un regard extérieur et des encouragements qui m'ont énormément touchés et qui m'ont permis d'avancer. Je souhaiterais également remercier Jean-Michel Marin. Mon goût pour les statistiques me vient en grande partie de toi. Tu as été un professeur qui m'a grandement inspiré, tes cours m'ont habité, ton enthousiasme et ton humour m'ont été très importants. Merci aussi à Jean-Noël Bacro d'avoir été d'une oreille attentive, d'avoir été un soutien et un garant du bon déroulement administratif de cette thèse. Merci à Bernadette Lacan, pour toutes ces discussions, pour toute ton aide extrêmement précieuse ! Tu m'as guidé dans tous mes problèmes administratifs, tu as pris beaucoup de temps pour moi et je t'en suis énormément reconnaissant ! Sophie Cazanave, tu m'as aussi beaucoup aidé pour des procédures administratives. Tu as à chaque fois pris du temps pour bien m'expliquer ce que j'ignorais, même quand tu étais très occupée. Pour mes problèmes administratifs, j'ai aussi pu compter sur Véronique Sals. Merci pour ces pauses, ces discussions, et ta vision, toujours pleine d'humanité, de gentillesse et de fraîcheur.

J'aimerais aussi remercier toutes ces personnes qui m'ont aidé pendant la thèse. Merci

en particulier à Johannie de m'avoir prêté sa maison en Lozère pour rédiger au calme, loin de la ville, pendant les derniers moments de la thèse. Merci parce que ce cadre m'a permis de prendre un recul important. C'était une période de travail très efficace, et c'est en partie grâce à ça, et donc à toi, que j'ai réussis à finir dans les temps cette thèse. Merci aussi à ceux qui ont relu la thèse à la recherche des fautes et coquilles. David, merci pour tes conseils anglophones. Lydia, merci pour ton œil de lynx et pour tout le reste. Merci à Morgane d'avoir relu à plusieurs étapes du manuscrit !

J'en profite pour exprimer ma plus profonde gratitude à Morgane. Tu m'as énormément aidé pendant ces derniers mois de thèse. Entre tes relectures de toute la thèse et ton soutien moral constant et sans faille. Tu m'as mis dans les meilleures conditions pour travailler pendant les moments les plus difficiles. Grâce à toi, j'ai pu travailler et écrire comme jamais. Tu m'as écouté, tu m'as motivé, tu m'as laissé travailler quand j'en avais besoin et tu me déconnectais de la thèse quand il le fallait. En particulier, les soirées salsa et les week end à la campagne m'ont permis de souffler et de revenir au travail de thèse avec une motivation incroyable. Tu m'as aidé à toutes les étapes de mon travail : pendant le travail de recherche, pendant la rédaction de thèse et même pour l'organisation de la soutenance. Sans toi, je n'aurais clairement pas aussi bien vécu cette fin de thèse.

Merci aussi à tous mes amis et à ma famille qui m'ont écouté et remonté le moral quand ça n'allait pas. Merci bien sûr à ma mère et à mon père pour leurs soutiens et à mes frères pour leur présence. Merci à Boris, Alex, grand Thomas, Laura, Guilhem, Jean, petit Thomas, Louise, Sandra, Elliot, Sonia, Zach, Lydia, François, Yann, Marina, Myriam, Micha, Arnaud, merci pour tout ! Et bien sûr, merci à mes camarades sportifs Paul, Alex et Tim !

Pendant, ces trois années de thèse, j'ai vécu avec les membres du labo de maths. Merci à tous pour ces échanges et ces bons moments. Merci en particulier à tout les doctorants dont la liste ne pourrait être exhaustive. Merci pour ces repas, ces pauses café, ces soutiens. On vit la même chose, on se comprend, et on peut râler !
Quoiqu'il en soit, pendant ces trois ans, j'ai aussi été dans le labo Mistea à SupAgro avec des gens d'une grande humanité. Merci à Pascal, Patrice, Brigitte, Béné, Nadine, Bertrand, Céline, aux Nicolas, Danaï, Hazaël, Philippe, Martine... Alex et Cheikh, j'aurais aimé que vous ne partiez pas. Et puis surtout, un très grand merci à Malika et Véro pour tout ces moments, ces discussions, votre écoute et votre soutien !

Bien entendu, cette thèse est un accomplissement que je n'aurais pas pu faire sans la présence d'une grande quantité de personnes. Même si ce n'est qu'au travers d'échanges anodins, vous m'avez épaulé et accompagné. Pendant trois ans, même des maillons parfois indiscernables peuvent être importants. Chacun à sa mesure, on apporte quelque chose et pour symboliser cela, je voudrais remercier mon plus grand fournisseur de nourriture pendant ces trois ans de thèse, mon pizzaiolo préféré : mon bonhomme de Show Pizza. Ce ne sont peut-être que de simples conversations, mais c'est parce qu'on sait qu'on peut compter dessus et qu'elles sont toujours aussi agréables ; qu'elles sont si importantes. Mettre bout à bout chacune de ces personnes du quotidien rend vertigineux notre place dans notre propre vie et me fait dire qu'il y une place pour lui dans ce travail de thèse.

Merci à tout ce que j'oublie honteusement de citer et à tout mes précieux maillons indiscernables.

Table des matières

1 Introduction **1**

2 La méthode Bliss **11**

2.1 Introduction 12

2.2 The Bliss Method 14

2.2.1 Reducing the Model 14

2.2.2 Model on a single Functional Covariate 15

2.2.3 Model Choice 17

2.2.4 Estimation of the Support 18

2.2.5 Estimation of the Coefficient Function 19

2.2.6 Model with Several Functional Covariates 21

2.2.7 Implementation 22

2.3 Simulation Study 24

2.3.1 Simulation Scheme for Datasets with One Functional Covariate . . 24

2.3.2 Performances Regarding Support Estimates 26

2.3.3 Performances Regarding the Coefficient Function 27

2.3.4 Tuning the Hyperparameters 32

2.3.5 Simulation Study for Two Functional Covariates 32

2.4 Application to the Black Périgord Truffle Dataset 34

2.5 Conclusion 40

2.6 Appendices 41

2.6.1 Theoretical Results 41
2.6.2 Details of the Implementations 44
2.6.3 Computational Time 46
2.6.4 Model Choice by using BIC 47

3 Construction d'une loi *a priori* informative 53

3.1 Introduction 53
3.2 *A priori* basé sur des pseudo-données 58
3.2.1 Modèle 60
3.2.2 Propriétés *a posteriori* 61
3.2.3 Calibration des poids 62
3.3 *A priori* basé sur une pénalisation 66
3.3.1 Choix de la distance 68
3.3.2 Calibration de τ 68
3.4 Implémentation 71
3.5 Résultats numériques 72
3.5.1 Application sur des données simulées 72
3.5.2 Application sur les données de truffes noires du Périgord 79
3.6 Discussion 86
3.7 Annexe 90
3.7.1 Distributions conditionnelles complètes 90
3.7.2 Implémentation 90

4 Consistance de la méthode Bliss 93

4.1 Introduction 93
4.2 Notations 95
4.3 Hypothèses 96
4.4 Résultat 97

4.5 Preuves et lemmes 98

4.6 Discussions 105

4.7 Annexes 107

4.7.1 Loi *a priori* et voisinage de β_0 107

5 Perspectives **109**

6 Annexe **113**

6.1 Exemple de mise en œuvre de la méthode Bliss 113

Table des figures

1.1 **Estimations de la fonction coefficient et de son support sur des données simulées.** *Pour le graphique de gauche, la courbe en pointillé représente la fonction coefficient utilisée pour simuler les données. La courbe noire (resp. cyan) en trait plein représente l'estimateur de la fonction coefficient avec la perte quadratique (resp. avec une nouvelle perte). La carte de couleurs du rouge au blanc est une représentation de la distribution a posteriori de la fonction coefficient. La couleur rouge (resp. blanche) est utilisée pour représenter une forte (resp. faible) valeur de la densité a posteriori. Les zones rouges (resp. blanches) correspondent aux zones où il y a de fortes (resp. faibles) probabilités a posteriori de retrouver la vraie fonction coefficient. Sur le second graphique, la courbe grise représente $\alpha(t)$, la probabilité a priori que la fonction coefficient soit non-nulle en t, et la courbe noire représente la probabilité de cet événement a posteriori. Lorsqu'on fixe $\gamma = 1/2$, l'estimateur du support est donné par les segments rouges.* 6

2.1 **The full Bayesian model.** *The coefficient function $\beta(t) = \sum_{k=1}^{K} b_k \mathbf{1}\{t \in \mathcal{I}_k\}/|\mathcal{I}_k|$ defines both a projection of the covariate functions $x_i(t)$ onto $\mathbb{R}^K$ by averaging the function over each interval $\mathcal{I}_k$ and a prediction $\widehat{y}_i$ which depends on the vector $b = (b_1, \ldots, b_K)$ and the intercept μ.* 16

2.2 **Coefficient functions for numerical illustrations.** *The black (resp. red and blue) curve corresponds to the coefficient function of Shape "Step function" (resp. "Smooth" and "Spiky").* 25

2.3 **Prior (in gray) and posterior (in black) probabilities of being in the support computed on Datasets 1 and 2.** *Bayes estimates of support using Theorem 2.1 with $\gamma = 1/2$ are given in red.* 25

2.4 **Estimates of the coefficient function on Dataset 4 ($r = 3$, $\zeta = 1$).** *For each plot, the black dotted line is the true coefficient function (Step function, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of $\beta(t)$ are represented using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate.* 28

2.5 **Estimates of the coefficient function on Dataset 13 ($r = 3$, $\zeta = 1$)).** *For each plot, the black dotted line is the true coefficient function (Smooth, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of $\beta(t)$ are represented using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate.* 29

2.6 **Estimates of the coefficient function on Dataset 25 ($r = 1$, $\zeta = 1$)).** *For each plot, the black dotted line is the true coefficient function (Spiky, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of $\beta(t)$ are represented using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate.* 30

2.7 **The coefficient functions $\beta_1(t)$ and $\beta_2(t)$ used to generate datasets in Section 2.3.5.** *The dark (resp. red) line represents $\beta_1(t)$ (resp. $\beta_2(t)$).* . 34

2.8 **Cross-covariance matrix between the curves $x_{i1}(\cdot)$ and $x_{i2}(\cdot)$ detailed in Section 2.3.5 for different values of c.** *Each point (t, t') represents the cross-covariance between $x_{i1}(t)$ and $x_{i2}(t')$. Red (resp. yellow) represents high (resp. low) cross-covariance.* 35

2.9 **Prior (in gray) and posterior (in black) probabilities of being in the support for $c = 0$ and for $c = 0.9$.** *Bayes estimates of support using Theorem 2.1 with $\gamma = 1/2$ are given in red.* 36

2.10 **Estimates of the coefficient functions for $c = 0$ and for $c = 0.9$.** *For each plot, the black dotted line is the true coefficient function, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate. The marginal posterior distributions of $\beta_\theta(t)$ are represented by using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities.* 37

2.11 **Rainfall of the Truffle datsaset.** Left:*Plot shows the rainfall for each year, colour-coded by their truffle yield.* Right:*Autocorrelation of the 13 observed rainfall covariates, with lag in number of ten-day periods.* 38

2.12 **Sensitivity of Bliss to the value of K on the truffle dataset.** Left: *Boxplot of the posterior distribution of the variance of the error, σ^2, compared to the variance of the output y (red dashed line).* Right: *Posterior probability $\alpha(t|\mathcal{D})$ for different values of K.* 39

2.13 **Coefficient functions with different numbers of intervals, used to simulate datasets.** 49

2.14 **The values of BIC for different values of the number of intervals of the true coefficient function.** 50

2.15 **The values of BIC for different levels of autocorrelation ζ.** *The left plot corresponds to datasets simulated for which the coefficient function is a step function with three intervals (see Figure 2.13). The right plot concerns simulated datasets for which the coefficient function is the Smooth function described in Section 3.1.* 50

2.16 **The values of BIC for different values of n.** 51

2.17 **The values of BIC for the truffle dataset described in Section 4.** . 51

3.1 **Distribution *a posteriori* de la fonction coefficient.** *Les graphiques (a), (b) et (c) donnent la distribution a posteriori sachant les pseudo-données, le graphique (d) sachant les données observées et les graphiques (e), (f) et (g) sachant les données observées et sachant les pseudo-données. Pour chaque graphique, la courbe noire est l'espérance a posteriori. Pour le graphique (a) (resp. (b) et (d)), la courbe noire en pointillé est la fonction coefficient utilisée pour générer les données $\mathcal{D}_1$ (resp. $\mathcal{D}_2$ et $\mathcal{D}_0$). Pour les graphiques (e), (f) et (g), la courbe cyan (resp. bleu) en pointillé est l'espérance sachant les données observées $\mathcal{D}_0$ (resp. les pseudo-données).* . . 74

3.2 **Exemples de distributions *a posteriori* pour différents niveaux de certitudes.** *Chacun des graphiques represente la distribution a posteriori de la fonction coefficient sachant $\mathcal{D}_0$ (données observées) et $\mathcal{D}_1$, $\mathcal{D}_2$ (pseudo-données). La courbe noire en trait plein est l'espérance a posteriori et la courbe cyan (resp. bleu) en pointillé est l'espérance sachant les données observées (resp. les pseudo-données).* 75

3.3 **Représentation de la connaissance des experts (simulés) en utilisant la procédure d'élicitation décrite en section 3.3.** *Le premier (resp. second) graphique donne la fonction $\bar{\beta}_E^s$ (resp. $\bar{g}_E$), la fonction signe moyenne des experts (resp. la certitude totale des experts).* 75

3.4 **Distributions *a posteriori* des termes $\frac{1}{2\sigma^2}$SCR et dist$(\boldsymbol{\beta^s, \bar{\beta}_E^s; \bar{g}_E})$ pour des données simulées lorsque τ est fixé par validation croisée.** *Le premier (resp. second) graphique représente la distribution a posteriori de $\frac{1}{2\sigma^2}SCR$ (resp. de $\mathrm{dist}(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$). La droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution.* . 77

3.5 **Résultats numériques pour différentes valeurs de τ.** *Le premier graphique donne les approximations numériques des utilités $u_{IS\text{-}LOO}(\tau)$ pour chaque valeur de τ dans $\boldsymbol{\tau}$. Le second graphique donne les approximations de l'utilité pour les valeurs de τ allant de 0 à $2 \times \tau_N$.* 77

3.6 **Résultats numériques pour différentes valeurs de τ lorsque la vraie fonction coefficient est de la forme *Smooth*.** *Le premier graphique donne les approximations numériques des utilités $u_{IS\text{-}LOO}(\tau)$ pour chaque valeur de τ dans $\boldsymbol{\tau}$. Le second graphique donne les approximations de l'utilité pour les valeurs de τ allant de 0 à $2 \times \tau_N$.* 78

3.7 **Distribution *a posteriori* de la fonction coefficient pour différents valeurs de τ, pour τ fixé par validation croisée.** *Le premier (resp. second) graphique représente la distribution a posteriori de la fonction coefficient quand $\tau = 0$ (resp. $\tau = 16.115$), ce qui correspond au cas où aucune connaissance d'expert n'est prise en compte. La courbe noire en trait plein est la fonction coefficient "Step function" utilisée pour générer les données. La courbe noire en pointillé est l'espérance a posteriori. La courbe cyan en pointillé pour le second graphique est l'espérance a posteriori quand $\tau = 0$.* 78

3.8 **Distributions *a posteriori* lorsque $\tau \sim \mathcal{E}(0)$.** *Le premier (resp. second) graphique représente la distribution a posteriori de la fonction coefficient (resp. de τ). Pour le premier graphique, la courbe noire en pointillé est l'espérance et celle en bleu est l'espérance si on fixe $\tau = 0$. Pour le second graphique, la droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution représentée.* 79

3.9 **Distributions *a posteriori* des termes $\frac{1}{2\sigma^2}$SCR et $\tau \times \mathrm{dist}(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$ pour des données simulées lorsque $\tau \sim \mathcal{E}(0)$.** *Le premier (resp. second) graphique représente la distribution a posteriori de $\frac{1}{2\sigma^2}SCR$ (resp. de $\tau \times \mathrm{dist}(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$). La droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution représentée.* . . 79

3.10 **Une partie du questionnaire donné aux experts. (1)** *Les experts doivent donner un scénario de précipitations, une production de truffes vraisemblable, et leur certitude en leur avis.* 81

3.11 **Une partie du questionnaire donné aux experts. (2)** *Pour un scénario de précipitations donné, les experts doivent donner une production de truffes vraisemblable et leur certitude en leur avis.* 82

3.12 **Résultats obtenus pour les données de truffes en prenant en compte les pseudo-données des experts.** *Le graphique* (a) *(resp.* (b)*) représente la distribution a posteriori de la fonction coefficient sachant les données observées (resp. les pseudo-données). Le graphique* (c) *montre la distribution a posteriori sachant les données observées et sachant les pseudo-données. Pour chaque graphique, la courbe noire en trait plein correspond à l'espérance associée à la distribution représentée. La courbe en pointillé cyan (resp. bleu) est l'espérance de la fonction coefficient sachant les données observées (resp. les pseudo-données).* 84

3.13 **Une partie du questionnaire donné aux experts. (3)** 85

3.14 **Résumé des avis des experts concernant le support et le signe de la fonction coefficient.** *Le premier graphique montre la fonction signe moyenne des experts* $\bar{\beta}_E^s$ *et le second donne leur certitude globale* $\bar{g}_E$. 85

3.15 **Résultats numériques pour les données de truffes quand** τ **est calibré par validation croisée bayésienne.** *Le premier graphique donne les valeurs du critère de validation croisée pour chaque valeur de* $\boldsymbol{\tau}$*, voir* (3.19)*. Le second graphique montre la distribution a posteriori de la fonction coefficient.* 86

3.16 **Distribution *a posteriori* de** $\frac{1}{2\sigma^2}$**SCR et de** $\tau \times \mathbf{dist^2}(\boldsymbol{\beta^s}, \bar{\boldsymbol{\beta}}_{\boldsymbol{E}}^{\boldsymbol{s}}; \bar{\boldsymbol{g}}_{\boldsymbol{E}})$ **pour les données de Pernes-Les-Fontaines (**τ **fixé par validation croisée bayésienne).** *Le premier (resp. second) graphique est un histogramme des valeurs de* $\frac{1}{2\sigma^2}SCR$ *(resp. de* $\tau \times \mathrm{dist}^2(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$*) calculé à partir de l'échantillon MCMC. Pour ces deux graphiques, la droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution représentée.* 86

3.17 **Résultats sur les données de truffes quand** $\tau \sim \mathcal{E}(0)$**.** *Le graphique* (a) *(resp.* (b)*) représente la distribution a posteriori de la fonction coefficient (resp.* τ*). La courbe noire (resp. cyan) en pointillé est l'espérance a posteriori quand* $\tau \sim \mathcal{E}(0)$ *(resp. quand* τ *est fixé à 0 : pas de pénalisation). Le graphique* (c) *(resp.* (d)*) représente la distribution a posteriori du terme de pénalisation* $\tau \times \mathrm{dist}^2(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$ *(resp.* $\frac{1}{2\sigma^2}SCR$*). Pour les graphiques* (a)*,*(b) *et* (c)*, la droite verticale rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution.* 87

6.1 **Représentation des courbes** $x_1(\cdot), \ldots, x_n(\cdot)$ **issues du jeu de données simulées.** 114

6.2 **Représentation des distributions *a posteriori* de la fonction coefficient et de son support.** *Le premier graphique est une représentation des probabilités a posteriori $\alpha(t|\mathcal{D})$, voir section 2.2.4 pour une description détaillée. Le second graphique est une représentation de la distribution a posteriori de la fonction coefficient. La couleur rouge (resp. blanche) est utilisée pour représenter une forte (resp. faible) valeur de la densité a posteriori.* 116

6.3 **Représentation de l'estimateur du support et des estimateurs de la fonction coefficient.** *Pour le premier graphique, l'estimateur du support est donné par les segments rouges. Pour le second graphique, l'estimation constante par morceaux de la fonction coefficient est donnée par la courbe noire en trait plein. L'estimateur lisse est donné par la courbe en pointillé. La courbe verte correspond à la vraie fonction coefficient utilisée pour générer les données.* 118

I

Introduction

Présentation générale

Les progrès technologiques de ces dernières décennies ont permis la collecte de grandes quantités de données. En particulier, les mesures de grandeurs climatiques comme la température ou les précipitations peuvent maintenant se faire de manière automatisée sur des pas de temps de plus en plus petits. Avant ces progrès technologiques, l'étude statistique considérait généralement une série temporelle de manière vectorielle. La fréquence d'observation devenant élevée, il est maintenant plus approprié de modéliser une série temporelle par une courbe qu'il convient de reconstruire à partir de mesures en différents instants. Avec une information suffisamment fine, on peut retrouver assez fidèlement cette courbe. On peut alors déterminer des caractéristiques à temps long (effet d'une tendance climatique) comme des caractéristiques à temps court (effet quotidien).
En agronomie, le développement des capteurs permet de mesurer le vécu de plantes durant toute une saison avec un pas de temps de l'ordre de la minute. Un des enjeux est alors d'établir des liens entre cette masse de données et des caractéristiques finales des cultures, comme un rendement ou le taux d'une molécule donnée dans les fruits. Dans le domaine de la trufficulture, une problématique importante est la compréhension des variations du rendement d'une truffière d'une année sur l'autre. On peut chercher à prédire cette quantité en prenant en compte des grandeurs climatiques comme les précipitations ayant vraisemblablement un impact sur le cycle de vie de la truffe. Cependant, au-delà de la prédiction, les agronomes cherchent à comprendre la réaction de la truffe aux conditions climatiques. La finalité peut être de guider l'agriculteur dans la gestion des cultures, ou dans d'autres contextes, de sélectionner une variété à cultiver en fonction du climat de la région.

D'un point de vue statistique, on cherche à expliquer une variable réelle à partir d'une ou plusieurs covariables fonctionnelles. Pour l'exemple de la truffe, les variations du climat x sont utilisées afin d'expliquer les variations des productions y. Un modèle statistique standard pour ce genre de problématique est le modèle de régression. Pour des

observations $y_1, \ldots, y_n$ et $x_1, \ldots, x_n$, une écriture générale de ce modèle est donnée par :

$$y_i | x_i \sim P_i, \quad \text{pour } i = 1, \ldots, n,$$

où P_i est une loi de probabilité qui dépend de la donnée x_i et qu'il reste à déterminer. L'estimation d'une loi de probabilité revient à en choisir une parmi un ensemble de lois de probabilité. Or, l'ensemble de toutes les lois de probabilité étant un ensemble complexe, le choix s'avère difficile et on se restreint généralement à un sous-ensemble de lois. Par soucis d'efficacité et de simplicité, on considère l'ensemble des lois normales $\mathcal{N}(r(x_i), \sigma^2)$ où $r(x_i)$ est une fonction affine de la courbe x_i :

$$r(x_i) = \mu + \int_0^1 x_i(t)\beta(t)\,\mathrm{d}t, \tag{1.1}$$

où 0 symbolise le début de la saison et 1 la fin. Ce modèle et plus généralement les données fonctionnelles ont été largement étudiés et popularisés par des ouvrages comme Ramsay and Silverman (1997) et Ferraty and Vieu (2006). On trouve dans la littérature différentes variantes de ce modèle, comme le modèle de régression fonctionnelle généralisé (Müller and Stadtmüller, 2005), le modèle mixte fonctionnel (Guo, 2002), le modèle de mélange de régression fonctionnelle (Yao et al., 2010), le modèle avec un terme d'interaction entre covariables fonctionnelles (Yang et al., 2013), le modèle pour lequel le domaine des covariables fonctionnelles est inconnu (Hall and Hooker, 2016), le modèle de régression fonctionnel avec une réponse fonctionnelle (Yao et al., 2005) ou encore pour des données fonctionnelles multidimensionnelles (Marx and Eilers, 2005). Pour une introduction plus exhaustive de ces modèles et des méthodes proposées pour les ajuster, on pourra consulter Müller (2005); Ullah and Finch (2013); Morris (2015); Reiss et al. (2016) et Wang et al. (2016).

Lorsqu'on cherche à comprendre la liaison entre x et y plutôt qu'à prédire un nouvel y, il faut principalement être capable d'interpréter la fonction β qui représente l'impact de x sur y. Il est pour cela nécessaire d'estimer β qui est une quantité de dimension infinie, ce qui est un problème statistique complexe. Les approches pour l'estimer consistent en général à réduire la dimension en décomposant β dans une base finie de fonctions. Plusieurs bases de fonctions sont envisageables et la dimension de la base est aussi à déterminer. Le coefficient fonctionnel se résume alors à un vecteur fini de coefficient dans la base et β s'estime en estimant ce vecteur. Cependant, il n'existe pas de choix canonique pour la base de fonctions ni pour la dimension de la base. Or suivant ces choix, l'estimation de β peut être sensiblement différente, ce qui rend difficile l'interprétation d'une estimation de β.

Une autre source d'erreur complique l'interprétation de la fonction coefficient. Il est connu dans un contexte de régression que la corrélation entre les covariables diminue la qualité de l'estimateur du coefficient de pente. Or, pour des données fonctionnelles, il y a par essence de grandes corrélations. En effet, la mesure d'une courbe en un instant t est hautement corrélée à la mesure à l'instant $t+h$, si h est petit. Pour pallier ce problème, une approche usuelle fournissant des estimateurs robustes consiste à minimiser un critère des moindres carrés pénalisé, comme la régression *Rigde* ou *Lasso*. En plus de prévenir les effets de corrélations dans les données, la régression *Lasso* effectue une sélection de variables et a donc de bonnes propriétés lorsque l'objectif est d'interpréter l'estimation. Dans le cadre du modèle de régression appliqué à des données fonctionnelles, les approches de pénalisation

ont été introduites pour régulariser l'estimateur de la fonction coefficient. En adaptant les idées de O'Sullivan (1986); Eilers and Marx (1996) concernant l'estimation de la fonction de régression, Marx and Eilers (1999); Cardot et al. (2003) introduisent une pénalité sur la dérivée seconde de la fonction coefficient pour lisser l'estimateur. Différentes bases de fonctions et différents ordres de dérivées ont été envisagés, mais malgré la robustesse des estimateurs proposés, rien ne garantit généralement qu'on soit capable d'interpréter l'estimation.

Pour comprendre comment on pourrait interpréter la fonction coefficient, notons que si $\beta(t)$ est nul sur une période T, on peut réécrire l'intégrale $\int_0^1 x_i(t)\beta(t)\,\mathrm{d}t$ sans faire intervenir les valeurs de $x_i(t)$ sur T. En ce sens, les valeurs de $x_i(t)$ pour $t \in T$ n'ont pas d'impact sur y_i du point de vue du modèle. Le support de la fonction coefficient est donc une caractéristique importante pour l'interprétation. De plus, selon le modèle, si $\beta(t)$ est positif (resp. négatif) sur une période, l'augmentation de la valeur de $x_i(t)$ sur cette période induit une augmentation (resp. diminution) de la valeur de y_i. On comprend aussi que plus la magnitude de $\beta(t)$ est élevée, plus l'impact (positif ou négatif) est important. Il est donc important de mettre en lumière les périodes pour lesquelles $\beta(t)$ est positif et celles pour lesquelles $\beta(t)$ est négatif. Cependant, si pour une période donnée, les valeurs de $\beta(t)$ fluctuent, l'interprétation n'est pas pour autant évidente. Or, les approches existantes fournissent généralement des estimations lisses de la fonction coefficient. Dans ce cas, on ne peut pas déterminer le support de β et les fluctuations de l'estimation sont des obstacles à l'interprétation. En ce sens, nous considérons qu'une estimation dont le support est clairement défini et qui est constante sur différentes périodes, est un bon candidat pour faciliter l'interprétation. Une fonction constante par morceaux et nulle sur plusieurs périodes vérifie ces propriétés. Dans la suite, nous nous concentrerons sur ces fonctions et nous considérerons ainsi une version simplifiée d'un modèle complexe. Quoiqu'il en soit, il existe d'autres problèmes qui rendent complexe l'interprétation de β. Du fait de la corrélation dans les données et des méthodes d'estimation, l'estimation de $\beta(t)$ sur une période dépend de l'estimation de $\beta(t)$ sur d'autres périodes. L'estimation de β sur une période est une information conditionnelle alors qu'une information jointe ou marginale serait plus appropriée pour interpréter l'impact de x sur y pour une période donnée.

Afin d'obtenir une estimation ayant une forme favorisant l'interprétation, James et al. (2009) introduisent des pénalités *Lasso* sur la fonction coefficient et sur une d^e dérivée. Du fait de la pénalisation *Lasso* sur β, l'estimation est parcimonieuse et la pénalité sur une dérivée impose une certaine régularité à l'estimation. Par exemple, en choisissant de pénaliser la dérivée première, on induit une parcimonie sur la dérivée première de l'estimation de β et on favorise donc une estimation constante par morceaux. Si on choisit de pénaliser la dérivée seconde, on favorise une estimation linéaire par morceaux. Pour appliquer cette méthode, il est nécessaire de calibrer l'intensité respective de chacune des pénalités *Lasso*, que les auteurs proposent de fixer par une validation croisée. Cette méthode est une des seules qui permettent de déterminer les périodes pour lesquelles la fonction coefficient est non-nulle. D'autres méthodes permettent d'accomplir cet objectif, en se concentrant sur l'estimation du support. Dans un contexte similaire au notre, Picheny et al. (2016) considèrent le modèle *Sliced Inverse Regression* (Li, 1991) étendu au cas de données fonctionnelles pour détecter des intervalles pertinents de la fonction coefficient. Lorsque la covariable fonctionnelle est multidimensionnelle, Park et al. (2016) proposent de segmenter le domaine en se basant sur la structure d'auto-corrélation des

fonctions $x_i(\cdot)$, puis de sélectionner les segments ayant les meilleures performances prédictives. Une autre méthode en deux étapes est proposée par Zhou et al. (2013) pour estimer le support et la fonction coefficient simultanément. La première étape consiste à estimer le support, puis la fonction coefficient est estimée à la seconde étape à partir de l'estimation du support.

Pour les agronomes, il est important d'estimer ces périodes mais aussi d'avoir un certain niveau de confiance en l'estimateur. Les précédentes méthodes sont des approches fréquentistes et le calcul de la loi de l'estimateur n'est pas évident, si bien qu'on n'obtient pas directement des intervalles de confiance pour l'estimateur de β.
De plus, les agronomes ont des connaissances permettant d'avoir une idée préliminaire de la fonction coefficient. Pour l'étude du rendement de la truffe, ils peuvent présumer de la forme de la fonction coefficient grâce aux connaissances de la croissance et du mode de reproduction de la truffe.
Les attentes des agronomes et leurs connaissances préliminaires suggèrent qu'une approche bayésienne est adaptée dans ce contexte. Pour ajuster le modèle de régression linéaire appliqué à des données fonctionnelles, plusieurs approches bayésiennes ont été envisagées, et sans être exhaustif on peut lister Brown et al. (2001); Wang et al. (2007); Crainiceanu and Goldsmith (2010); Goldsmith et al. (2011). Les approches proposées considèrent généralement une loi *a priori* sur les coefficients de β dans une base de fonctions, alors que d'autres introduisent une loi *a priori* sur un espace de fonctions, voir par exemple Kang et al. (2016). A notre connaissance, aucune ne s'intéresse spécifiquement au support de la fonction coefficient ou à la prise en compte de connaissances préliminaires.

Dans un cadre plus général, il existe un large éventail de méthodes pour prendre en compte des connaissances préliminaires dans un cadre bayésien. Pour une synthèse de ces méthodes et leur limitations, on pourra consulter les ouvrages Cooke (1991); Meyer and Booker (2001); Ouchi (2004); Kynn (2005); O'Hagan et al. (2006); Kynn (2008); Low Choy (2012). Bien que de nombreuses méthodes existent, il est en général difficile de les adapter à un autre contexte. En effet, la plupart sont dépendantes du modèle étudié mais aussi du type de connaissances préliminaires disponibles (voir Kadane and Wolfson, 1998). Cependant, des méthodes plus générales existent, notamment pour le modèle de régression linéaire, comme Albert et al. (2012). Quoiqu'il en soit, aucune méthode ne s'intéresse au modèle de régression appliqué aux données fonctionnelles. De plus, le type de connaissances préliminaires des agronomes concernant le lien entre les précipitations et la production de truffes est spécifique au modèle de régression linéaire fonctionnel.

Pour répondre à ces problématiques, notre contribution principale dans cette thèse est de proposer la méthode *Bayesian functional Linear regression with Sparse Step function* (Bliss) pour estimer la fonction coefficient par une fonction constante par morceaux et pour estimer son support. De plus, pour inclure dans le modèle l'avis des agronomes, nous proposons des manières alternatives de construire la distribution *a priori*. Enfin, nous étudions le comportement asymptotique de la méthode proposée.

La méthode Bliss

Le modèle de régression linéaire fonctionnel que nous proposons est basé sur la restriction aux fonctions coefficient constantes par morceaux. Pour inclure cette contrainte dans le modèle, nous proposons de construire une loi *a priori* qui prenne en compte uniquement les fonctions en escalier de la forme :

$$\beta(t) = \sum_{k=1}^{K} \frac{b_k}{|\mathcal{I}_k|} \mathbf{1}_{\mathcal{I}_k}(t), \tag{1.2}$$

où les b_k sont des réels, les $\mathcal{I}_k$ sont des intervalles de $[0,1]$ et K est un hyperparamètre à déterminer. En prenant en compte la contrainte (1.2), le modèle de régression linéaire fonctionnel se réécrit comme un modèle de régression linéaire multiple où le *design* dépend des intervalles $\mathcal{I}_k$:

$$r(x_i) = \mu + \sum_{k=1}^{K} b_k \, x_i(\mathcal{I}_k), \quad \text{où } x_i(\mathcal{I}_k) = \frac{1}{|\mathcal{I}_k|} \int_{\mathcal{I}_k} x_i(t) \mathrm{d}t. \tag{1.3}$$

Nous définissons une paramétrisation des intervalles $\mathcal{I}_k$ en fonction des milieux $m_k \in [0,1]$ et des demi-longueurs $\ell_k > 0$. La loi *a priori* des ℓ_k favorise les valeurs proches de 0 si bien que l'union des K intervalles ne recouvre pas $[0,1]$. Nous considérons ainsi une fonction coefficient *a priori* parcimonieuse.

De la distribution *a posteriori*, nous déduisons plusieurs estimateurs bayésiens : un estimateur du support et deux estimateurs de la fonction coefficient aux propriétés différentes. Pour déterminer un estimateur du support S, on distingue deux types d'erreurs possibles : soit on estime à tort que $t \in S$, soit on estime à tort que $t \notin S$. Naturellement, on considère qu'un estimateur $\hat{S}$ doit réaliser un compromis entre ces deux types d'erreurs, ce qui revient à minimiser la perte

$$L_\gamma(S, \hat{S}) = \gamma \int_0^1 \mathbf{1}\{t \in \hat{S} \setminus S\} \, \mathrm{d}t + (1-\gamma) \int_0^1 \mathbf{1}\{t \in S \setminus \hat{S}\} \, \mathrm{d}t. \tag{1.4}$$

Le paramètre $\gamma \in [0,1]$ calibre l'importance relative d'une erreur par rapport à l'autre et sans information supplémentaire, on peut fixer $\gamma = 1/2$. On montre avec le théorème 2.1 que minimiser l'espérance *a posteriori* de cette perte revient à déterminer les valeurs de t pour lesquelles $\alpha(t|\mathcal{D}) \geq \gamma$, où $\alpha(t|\mathcal{D})$ est la probabilité que $\beta(t)$ soit non-nul, sachant les données $\mathcal{D}$.
Pour estimer la fonction coefficient, une première approche est de considérer l'estimateur bayésien classique, basé sur la perte quadratique. Cependant, cet estimateur n'est pas une fonction en escalier puisque l'ensemble des fonctions en escalier qui vérifient (1.2) n'est pas convexe. Une étude de simulation illustre qu'il reconstruit cependant correctement la vraie fonction coefficient quelle que soit sa régularité.
Afin d'obtenir une estimation constante par morceaux, nous proposons de considérer une nouvelle fonction de perte, qui attribue un coût infini aux fonctions qui ne respectent pas la contrainte (1.2). La proposition 2.3 montre que cette perte définit un estimateur bayésien qui correspond à la projection de l'estimateur précédent sur l'ensemble des fonctions vérifiant la contrainte du modèle.

Finalement, on peut résumer les caractéristiques principales de la méthode Bliss par deux graphiques donnés dans la figure 1.1. Le premier illustre les deux estimateurs de la

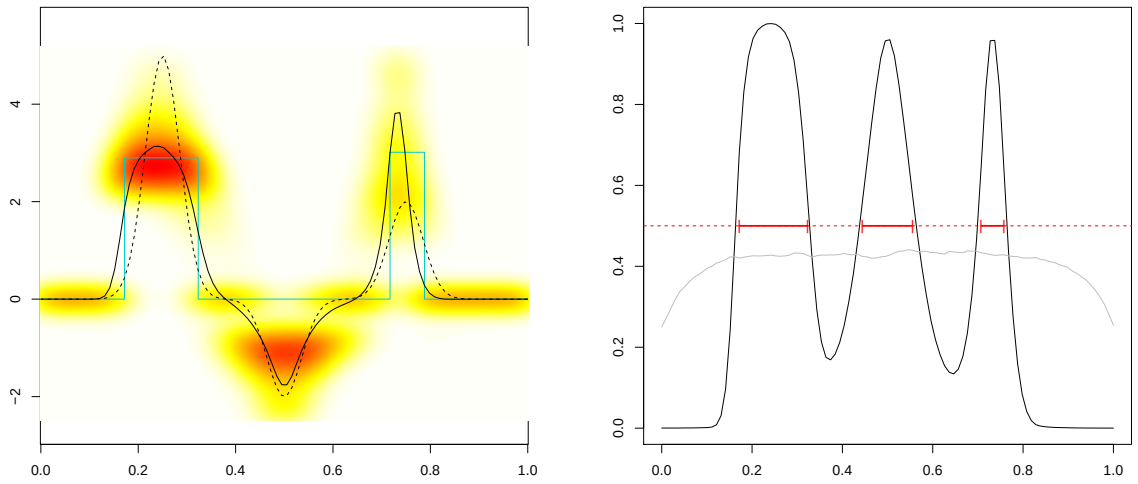


Figure 1.1: Estimations de la fonction coefficient et de son support sur des données simulées. *Pour le graphique de gauche, la courbe en pointillé représente la fonction coefficient utilisée pour simuler les données. La courbe noire (resp. cyan) en trait plein représente l'estimateur de la fonction coefficient avec la perte quadratique (resp. avec une nouvelle perte). La carte de couleurs du rouge au blanc est une représentation de la distribution a posteriori de la fonction coefficient. La couleur rouge (resp. blanche) est utilisée pour représenter une forte (resp. faible) valeur de la densité a posteriori. Les zones rouges (resp. blanches) correspondent aux zones où il y a de fortes (resp. faibles) probabilités a posteriori de retrouver la vraie fonction coefficient. Sur le second graphique, la courbe grise représente $\alpha(t)$, la probabilité a priori que la fonction coefficient soit non-nulle en t, et la courbe noire représente la probabilité de cet événement a posteriori. Lorsqu'on fixe $\gamma = 1/2$, l'estimateur du support est donné par les segments rouges.*

fonction coefficient et le second, l'estimateur du support. Pour ces graphiques, les estimations sont obtenues pour des données simulées, générées à partir de la fonction coefficient donnée sur le premier graphique par la courbe noire en pointillé. Pour chacun des estimateurs proposés, la distribution *a posteriori* permet de déterminer directement la crédibilité de ces résultats. Avec ces différents estimateurs, notre méthode permet de répondre à plusieurs problématiques : estimer avec précision la fonction coefficient, fournir une estimation qui facilite l'interprétation ou détecter le support de la fonction coefficient. Nous avons, de plus, implémenté cette méthode dans un langage de programmation efficace et nous nous montrons que le temps de calculs pour appliquer la méthode est raisonnable.

Cependant, un des facteurs limitant de l'approche proposée est de devoir choisir un hyperparamètre K, ayant une influence majeure sur la distribution *a posteriori*. Pour le calibrer, nous proposons d'utiliser le critère BIC dont les bonnes performances sont illustrées par une étude de simulation.
Une autre limitation vient de la paramétrisation (1.2). Nous autorisons les intervalles à se chevaucher, cependant si deux intervalles sont identiques, la matrice de Gram du *design* est singulière. Du fait de problèmes numériques, cette matrice est mal conditionnée dès lors que deux intervalles sont presque identiques. Pour éviter ce problème, les hyperparamètres b_k suivent une loi *a priori* inspirée de celui de *Ridge-Zellner*. Cette distribution *a priori* fait intervenir un terme de pénalisation qu'il est alors nécessaire de calibrer.
En ce qui concerne le calcul de l'estimation constante par morceaux, un algorithme d'op-

timisation est nécessaire pour minimiser l'espérance *a posteriori* de la perte.

Nous appliquons la méthode proposée pour déterminer comment l'évolution des précipitations dans le temps influence la production de truffes noires du Périgord. Nous mettons en lumière une période avec un impact négatif au printemps et une période avec un impact positif en été, ce qui est pertinent du point de vue des trufficulteurs et des biologistes. Pour cette étude, les agronomes ont des connaissances préliminaires suggérant que la truffe est plus sensible aux niveaux de précipitations pendant certaines périodes. Afin d'intégrer ces connaissances utiles dans le modèle Bliss, nous construisons la loi *a priori* à partir d'informations données par les experts.

Construction d'une loi *a priori* informative

Nous proposons deux méthodes pour construire la distribution *a priori* en prenant en compte ces connaissances. La première méthode consiste à éliciter l'avis de E experts en leur demandant à chacun de construire des pseudo-données. Les pseudo-données $y^e = (y_1^e, \ldots, y_{n_e}^e)$ d'un expert e, associées à des courbes $x^e = (x_1^e, \ldots, x_{n_e}^e)$, représentent l'avis de l'expert concernant l'impact de la covariable fonctionnelle sur la variable réponse. Nous modélisons conjointement les données observées $y = (y_1, \ldots, y_n)$ et les pseudo-données $y^1, \ldots, y^E$ avec le modèle Bliss. Cependant, pour prendre en compte la différence entre ces deux types de données, nous proposons une approche s'inspirant à la fois de la régression par vraisemblance pondérée (Hu and Zidek, 1995) et de l'approche *Power prior* (Ibrahim and Chen, 2000). La régression par vraisemblance pondérée consiste à attribuer un poids à chacune des données afin de contrôler son impact dans l'ajustement du modèle. La seconde approche considère comme loi *a priori*, une distribution *a posteriori* sachant des données historiques. Le poids des données historiques est relativisé par un paramètre à calibrer. L'approche que nous proposons a pour but de contrôler l'influence des pseudo-données en considérant la vraisemblance pondérée

$$L(y, y^1 \ldots, y^E|\theta; w) = \prod_{i=1}^{n} p(y_i|\theta) \times \prod_{e=1}^{E} \prod_{i=1}^{n_e} p(y_i^e|\theta)^{w_i^e}, \tag{1.5}$$

où les w_i^e sont des poids à calibrer et $p(\cdot)$ est la vraisemblance du modèle Bliss. Ce qui revient, de manière équivalente, à considérer un modèle pour les données observées avec comme loi *a priori*, la loi *a posteriori* sachant les pseudo-données :

$$\pi(\theta|y^1, \ldots, y^E; w) = \pi(\theta) \times \prod_{e=1}^{E} \prod_{i=1}^{n_e} \pi(y_i^e|\theta)^{w_i^e}. \tag{1.6}$$

La distribution *a posteriori* de ce modèle se décompose comme une partie apportée par les données observées et une partie apportée par les pseudo-données. De cette décomposition découle deux propriétés importantes qui permettent d'interpréter les poids w_i^e. Premièrement, si $w_i^e = 0$, la distribution *a posteriori* ne dépend pas de la pseudo-donnée y_i^e. A l'opposé, si $w_i^e = 1$, la pseudo-donnée aura autant d'importance qu'une donnée observée. La seconde propriété permet d'interpréter l'estimateur de la fonction coefficient comme une moyenne entre un estimateur sachant les données observées et un estimateur sachant les pseudo-données, pondérée par les poids w_i^e. De plus, on déduit de ces propriétés que les poids w_i^e ont une importance majeure dans l'inférence. Pour les calibrer, nous proposons

de nous appuyer sur la confiance des experts en leur avis, que nous relativisons selon deux principes. Premièrement, nous prenons en compte la corrélation entre les experts afin de ne pas prendre en compte d'informations redondantes. Ensuite, nous renormalisons les w_i^e afin que le poids des pseudo-données n'excède pas le poids des données observées. La première approche que nous proposons ici est assez générique. En effet, il suffit que les experts puissent construire des pseudo-données et que le statisticien détermine un calibration appropriée des poids.

La seconde méthode que nous proposons est plus spécifique au modèle Bliss puisque nous modélisons l'avis des experts concernant la fonction coefficient. Nous prenons en compte des avis qui se décomposent en deux parties : 1) la connaissance du support de la fonction coefficient et 2) la connaissance de son signe sur ce support. Modéliser cet avis revient à considérer la fonction signe $\beta^s(\cdot)$:

$$\beta^s(t) = \mathbf{1}\left\{\beta(t) > 0\right\} - \mathbf{1}\left\{\beta(t) < 0\right\}, \tag{1.7}$$

et on traduit les avis des experts en des fonctions signes $\beta_e^s(\cdot)$. D'un point de vue modélisation, pour que l'inférence tienne compte de l'avis des experts, nous pénalisons les fonctions coefficients qui ne sont pas *en accord* avec ces avis. Pour cela, nous introduisons une nouvelle distribution *a priori* qui dépend d'un hyperparamètre τ et de la distance entre la fonction signe β_s et les fonctions signes des experts β_e^s. Il en découle la distribution *a posteriori* :

$$\pi(b, \mathcal{I}|y, \mu, \sigma^2; \tau) \propto \exp\left\{-\left(\frac{1}{2\sigma^2}\mathrm{SCR} + \tau \sum_{e=1}^{E} \mathrm{dist}^2\left(\beta^s, \beta_e^s; g_e\right)\right)\right\} \times \pi_0(b, \mathcal{I}|\sigma^2), \tag{1.8}$$

où π_0 est la loi *a priori* du modèle Bliss. Pour calibrer la valeur de τ, nous proposons deux approches. La première consiste à choisir τ parmi un ensemble de candidats en utilisant une procédure bayésienne de validation croisée. Pour appliquer cette procédure, on définit l'utilité d'une valeur de τ et on choisit τ qui maximise l'utilité. La seconde option est de considérer τ aléatoire et de lui attribuer une distribution *a priori*. Par défaut, on choisit une loi conjuguée et τ suit alors *a priori* une loi exponentielle.

Consistance de la méthode Bliss

Pour valider la méthode Bliss, nous discutons de son comportement asymptotique. En particulier, nous étudions la consistante de la distribution *a posteriori*, c'est-à-dire si la probabilité *a posteriori* du voisinage du *vrai* paramètre $\theta_* = (\mu_*, \beta_*, \sigma_*^2)$ tend vers 1 lorsque la taille d'échantillon augmente, et ce quel que soit ce voisinage. Dans notre cas, il y a principalement deux difficultés pour démontrer cette consistance. Non seulement, dans un contexte de régression, les données ne sont pas identiquement distribuées mais, de plus, le modèle Bliss est mal spécifié. En effet, le modèle suppose que la fonction coefficient est une fonction en escalier, or la *vraie* fonction β_* n'est sûrement pas en escalier avec K plateaux. La distribution *a posteriori* ne peut donc pas se concentrer autour du *vrai* paramètre. Dans ce cas, les résultats de consistance classiques établissent la concentration autour d'un paramètre $\theta_0 = (\mu_0, \beta_0, \sigma_0^2)$, dit le *pseudo-vrai* paramètre. De manière générale, l'outil le plus utilisé pour montrer la consistance est le théorème de Schwartz (1965). Lorsque le modèle est mal spécifié, des hypothèses plus complexes sont

nécessaires afin d'obtenir ce théorème, voir Kleijn and van der Vaart (2006). Dans des contextes de régression, Amewou-Atisso et al. (2003); Choi and Schervish (2004); Ghosal et al. (2007) étendent ce théorème au cas de données non-identiquement distribuées.

Plutôt que de vérifier des hypothèses, complexes dans le cas du modèle Bliss, nous proposons des hypothèses spécifiques à notre modèle pour écrire une démonstration efficace qui s'apparente à celle de Wald (1949). Parmi les hypothèses nécessaires, nous avons besoin de deux hypothèses sur le *design*. La première, raisonnable en pratique, suppose que le *supremum* de $|x_i(t)|$ ne dépasse pas un certain seuil, quel que soient $i \geq 1$ et $t \in [0,1]$. La seconde hypothèse impose une certaine régularité au *design*. Plus précisément, elle suppose que la fonction moyenne $t \mapsto n^{-1}\sum_{i=1}^n x_i(t)$ et la fonction $(t,t') \mapsto n^{-1}\sum_{i=1}^n x_i(t)\times x_i(t')$ convergent respectivement simplement vers $e(\cdot) \in L^2([0,1])$ et vers $c(\cdot,\cdot) \in L^2([0,1]\times[0,1])$. De plus, étant donné la contrainte (1.2) du modèle Bliss, nous avons besoin de l'hypothèse suivante, qui suppose l'existence de la solution d'un problème de minimisation dans la démonstration du théorème.

Hypothèse. *Il existe une unique fonction* $\beta_0(t) = \sum_{k=1}^K b_{0k}\mathbf{1}_{\mathcal{I}_{0k}}(t)$ *qui minimise*

$$F(\beta) = \iint (\beta_* - \beta)(t) \times (\beta_* - \beta)(t')\Big[c(t,t') - e(t)e(t')\Big]\,\mathrm{d}t\,\mathrm{d}t'.$$

Nous considérons une dernière hypothèse selon laquelle l'espace paramétrique est compact. Cette hypothèse classique permet d'avoir un cadre simplifié pour établir le résultat suivant.

Théorème. *Soit U le complémentaire d'un voisinage de θ_0. Sous les hypothèses définies précédemment, la probabilité a posteriori de U, $\Pi_n(U)$, tend $P_{\theta_*}^\infty$-presque sûrement vers 0, quand $n \to +\infty$.*

En plus de ce résultat, nous montrons que θ_0 correspond à la projection de θ_* au sens de la divergence de Kullback-Leibler et nous donnons son expression à partir du *design* et de θ_*. En particulier, nous quantifions l'erreur d'estimation sur μ_* et sur σ_*^2 en fonction de l'erreur commise sur β_*. D'un point de vue fréquentiste, le théorème de consistance donne la convergence des estimateurs proposés. Pour une approche bayésienne, ce type de résultat atteste qu'avec deux distributions *a priori* différentes vérifiant les hypothèses du théorème, l'inférence donnera le même résultat. En ce sens, ce résultat valide la méthode Bliss ainsi que les variantes construisant une loi *a priori* à partir des avis d'experts. De plus, ce résultat permet de contourner le problème du choix de K. En effet, si la *vraie* fonction coefficient est une fonction en escalier, alors il suffit de prendre K assez grand pour assurer la consistance de la distribution *a posteriori*. Au contraire, si on considère que ce n'est pas une fonction en escalier, alors il n'existe pas de bonne valeur de K et on se contentera de la plus grande valeur de K possible en pratique.

Plan de la thèse

Le reste de ce manuscrit se décompose en trois chapitres principaux, ainsi que d'un chapitre de perspectives et d'une annexe. Le chapitre 2 est le cœur de la thèse, dans lequel nous introduisons en détails le modèle Bliss qui sera étudié dans l'ensemble de la

thèse. Ce chapitre sera l'occasion d'illustrer les performances des estimateurs proposés avec une étude sur des données simulées. Cette méthode est étendue, dans le chapitre 3, pour prendre en compte des informations préliminaires d'experts concernant la fonction coefficient. Le chapitre 4 est dédié à l'étude du comportement asymptotique de la méthode Bliss. A partir des travaux présentés dans ces trois chapitres, émergent des perspectives de recherche dont nous discuterons dans le chapitre 5. Finalement, dans l'annexe 6, nous illustrons l'implémentation de la méthode Bliss avec un exemple détaillé sur des données simulées.

II

La méthode Bliss

Un outil fondamental en statistique pour déterminer l'impact d'une covariable fonctionnelle sur une variable réelle, est le modèle de régression linéaire fonctionnel, dont le terme central est la fonction coefficient. Dans ce chapitre, nous proposons une approche bayésienne qui s'intéresse en particulier au support de cette fonction coefficient. Pour cela, nous considérons une décomposition parcimonieuse et adaptative de la fonction coefficient en une fonction en escalier. Le modèle bayésien que nous proposons, nommé *Bayesian functional Linear regression with Sparse Step functions* (Bliss), est basé sur cette décomposition. Notre objectif est d'estimer les périodes qui influencent le plus la variable réelle. Nous construisons un estimateur bayésien du support en considérant une nouvelle fonction de perte, ainsi que deux estimateurs bayésiens de la fonction coefficient, un lisse et un constant par morceaux. Les performances de la méthode proposée sont étudiées sur des données simulées. En particulier, nous évaluons la robustesse des estimateurs par rapport au choix des hyperparamètres du modèle. Nous comparons aussi les estimateurs proposés à des méthodes concurrentes sur des jeux de données ayant des caractéristiques différentes. La méthode Bliss est ensuite utilisée pour étudier l'influence des précipitations sur la production de truffes noires du Périgord.

Dans le reste du chapitre, nous donnons l'article soumis dans une revue anglophone qui introduit la méthode Bliss et les estimateurs proposés. L'annexe de ce chapitre comprend un document supplémentaire attaché à l'article. Dans ce document, nous discutons notamment du temps de calcul de la méthode Bliss. De plus, nous illustrons les performances du critère BIC pour le choix de K. Pour ce qui concerne l'implémentation de la méthode, les développements de ce chapitre sont illustrés dans le chapitre annexe 6 dans lequel nous présentons une mise en œuvre de la méthode sur des données simulées.

Contents

2.1 Introduction **12**
2.2 The Bliss Method **14**
2.3 Simulation Study **24**
2.4 Application to the Black Périgord Truffle Dataset **34**
2.5 Conclusion **40**
2.6 Appendices **41**

2.1 Introduction

Consider that one wants to explain the final outcome y of a process along time (for instance the amount of some agricultural production) thanks to what happened during the whole history (for instance, the rainfall history, or temperature history). Among the statistical learning methods, functional linear models (Ramsay and Silverman, 2005) aim at predicting a scalar y based on covariates $x_1(t), x_2(t), \ldots, x_q(t)$ lying in a functional space, $L^2(\mathcal{T})$ say, where $\mathcal{T}$ is an interval of $\mathbb{R}$. If $x_{q+1}, \ldots, x_u$ are additional scalar covariates, the outcome y is predicted linearly with

$$\widehat{y} = \mu + \int_{\mathcal{T}} \beta_1(t)x_1(t)\mathrm{d}t + \cdots + \int_{\mathcal{T}} \beta_q(t)x_q(t)\mathrm{d}t + \beta_{q+1}x_{q+1} + \cdots + \beta_u x_u, \tag{2.1}$$

where μ is the intercept, $\beta_1(t), \ldots, \beta_q(t)$ the coefficient functions, and $\beta_{q+1}, \ldots, \beta_u$ the other (scalar) coefficients. Note that nonlinear models have also been considered, see for instance Ferraty and Vieu (2006). In the linear framework, the functional covariates $x_j(t)$ and the unknown coefficient functions $\beta_j(t)$ lie in the $L^2(\mathcal{T})$ functional space, thus we face a nonparametric problem. Standard methods (Ramsay and Silverman, 2005) for estimating the $\beta_j(t)$'s, $1 \leq j \leq q$, are based on the expansion onto a given basis of $L^2(\mathcal{T})$ and the minimization of a penalized criterion to avoid overfitting, see for instance Cardot et al. (2003). The choice of the given basis is a main feature of these approaches and several choices have been considered as, for example, data-driven basis (see Cardot et al., 1999, Yuan and Cai, 2010 and Zhu et al., 2014) or fixed standard basis (see among others Crambes et al., 2009; Zhao et al., 2012). Bayesian functional regression methods mainly use Gaussian process prior (Behseta et al., 2005; Shi et al., 2007; Yang et al., 2017), Dirichlet process (Ray and Mallick, 2006; Petrone et al., 2009; Rodríguez et al., 2009; Yang et al., 2016) or put a prior distribution on the coefficient of the basis expansion (Brown et al., 2001; Crainiceanu et al., 2005; Crainiceanu and Goldsmith, 2010; Goldsmith et al., 2011; Montagna et al., 2012). Setting a hierarchical model or using a mixed model is a way to include prior knowledge in the model (Goldsmith et al., 2010, 2011). For a comprehensive scan of the methodology, see Morris (2015) and Reiss et al. (2016).

An issue which arises naturally in many applied contexts is the detection of periods of time which influence the final outcome y the most. Note that each integral in (2.1) is a weighted average of the whole trajectory of $x_j(t)$, and does not identify any specific

impact of specific periods of the process. These time periods might vary from one covariate to another. For instance, in agricultural science, the final outcome may depend on the amount of rainfall during a given period (e.g., to prevent rotting), and the temperature during another (e.g., to prevent freezing). Standard methods do not answer the above question, namely to recover the support of the coefficient functions $\beta_j(t)$ with the noticeable exception of Picheny et al. (2016).

Unlike the scalar-on-image models, we focus here on one-dimensional functional covariates. When $\mathcal{T}$ is not a one dimensional space, the problem becomes much more complex. The functional covariates and the coefficient functions are all discretized, e.g. via the pixels of the images, see Goldsmith et al. (2014); Li et al. (2015); Kang et al. (2016). In these two- or three-dimensional problems, because of the curse of dimensionality, the points which are included in the support of the coefficient functions follow a parametric distribution, namely an Ising model. One important problem solved by these authors is the sensitivity of the parameter estimate of the Ising model in the neighborhood of the phase transition.

When $\mathcal{T}$ is a one dimensional space, we can build nonparametric estimates. In this vein, using the L^1-penalty to achieve parsimony, the Flirti method of James et al. (2009) obtains an estimate of the $\beta_j(t)$'s assuming they are sparse functions with sparse derivatives. Nevertheless Flirti is difficult to calibrate: its numerical results depend heavily on tuning parameters. In our experience, Flirti's estimate is so sensitive to the values of the tuning parameters that we can miss the range of good values with cross-validation. The authors propose to rely on cross-validation to set these tuning parameters. But, by definition, cross-validation assesses the predictive performance of a model, see Arlot and Celisse (2010) and the many references therein. Of course, optimizing the performance regarding the prediction of y does not provide any guarantee regarding the support estimate. Zhou et al. (2013) propose a two-stage method to estimate the coefficient function. Beforehand, $\beta(t)$ is expanded onto a B-spline basis to reduce the dimension of the model. The first stage estimates the coefficients of the truncated expansion onto the basis using a lasso method to find the null intervals. Then, the second stage refines the estimation of the null intervals and estimates the magnitude of $\beta(t)$ for the rest of the support. Another approach to obtain parsimony is to rely on Fused lasso (Tibshirani et al., 2005): if we discretize the covariate functions and the coefficient function as described in James et al. (2009), the penalization of Fused lasso induces parsimony in the coefficients, but, once again the calibration of the penalization is performed using cross-validation which targets predictive perfomance rather than the accuracy of the support estimate.

In this paper, we propose Bayesian estimates of both the supports and the coefficient functions $\beta_j(t)$. To keep the dimension of the parameter as low as possible, we stay with the simplest and the most parsimonious shape of the coefficient function over its support. Hence, conditionally on the support, we consider the coefficient functions $\beta_j(t)$ to be step functions (piecewise constant functions can be described with a minimal number of parameters). We can decompose any step function $\beta(t)$ as follows:

$$\beta(t) = \sum_{k=1}^{K} b_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}\{t \in \mathcal{I}_k\}$$

where $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are intervals of $\mathcal{T}$, $|\mathcal{I}_k|$ is the length of the interval and b_k are the coefficients of the expansion. The support is the union of all $\mathcal{I}_k$ if the coefficients b_k are non null. A period of time which does not influence the outcome will be outside the

support. The above model has another advantage: step functions change values abruptly from 0 to a non null value. Hence their supports are relatively clear. On the contrary, if we have at our disposal a smooth estimate of a coefficient function $\beta_j(t)$ in the model given by (2.1), the support of the estimate is the whole $\mathcal{T}$ and we have to find regions where the estimate is not significantly different from 0. Moreover, with a full Bayesian procedure, we can evaluate the uncertainty of the estimates of the support and the values of the coefficient functions.

This paper is organized as follows. Section 2.2 presents the Bayesian modelling, including the prior distribution in 2.2.2, the support estimate in 2.2.4 and the coefficient function estimates in 2.2.5. Section 2.3 is devoted to the study of numerical results on synthetic data, with comparison to other methods and sensibility to the tuning of the hyperparameters of the prior. Section 2.4 gives details of the results of Bliss on a dataset concerning the influence of rainfall on the growth of the black Périgord truffle.

2.2 The Bliss Method

We present the hierarchical Bayesian model in Section 2.2.2 on a single functional covariate, the Bayes estimate of the support in Section 2.2.4 and two Bayes estimates of the coefficient function in Section 2.2.5. Section 2.2.6 describes the Bayesian model on several functional covariates. The implementation and visualization details are given at the end of this section.

2.2.1 Reducing the Model

Assume we have observed n independent replicates y_i $(1 \leq i \leq n)$ of the outcome, explained with the functional covariates $x_{ij}(t)$ $(1 \leq i \leq n,\ 1 \leq j \leq q)$ and the scalar covariates x_{ij} $(1 \leq i \leq n,\ q+1 \leq j \leq u)$. The whole dataset will be denoted $\mathcal{D}$ in what follows. Let us denote by $x_i = \{x_{i1}(t), \ldots, x_{iq}(t), x_{i,q+1}, x_{iu}\}$ the set of all covariates for replicate i, and by θ the set of all parameters, namely $\{\beta_1(t), \ldots, \beta_q(t), \beta_{q+1}, \ldots, \beta_u, \mu, \sigma^2\}$, where σ^2 is a variance parameter. We resort to the Gaussian likelihood defined as

$$y_i | x_i, \theta \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + \sum_{j=1}^{q} \int_{\mathcal{T}} \beta_j(t) x_{ij}(t) \mathrm{d}t + \sum_{j=q+1}^{u} \beta_j x_{ij},\ \sigma^2\right), \qquad i = 1, \ldots, n. \tag{2.2}$$

If we set a prior on the parameter θ which includes all $\beta_j(t)$, β_j, μ and σ^2, we can recover the full posterior from the following conditional distributions (both theoretically and practically with a Gibbs sampler) :

$$\beta_j(t), \mu, \sigma^2 \,|\, \mathcal{D}, \beta_{-j}$$
$$\beta_j, \mu, \sigma^2 \,|\, \mathcal{D}, \beta_{-j}$$

where β_{-j} represents the set of β-parameters except β_j or $\beta_j(t)$. Hence we can reduce the problem to a single functional covariate and no scalar covariate. The model we have to study becomes

$$y_i | x_i(t), \mu, \beta(t), \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + \int_{\mathcal{T}} \beta(t) x_i(t) \mathrm{d}t\,,\ \sigma^2\right), \qquad i = 1, \ldots, n, \tag{2.3}$$

with a single functional covariate $x_i(t)$.

2.2.2 Model on a single Functional Covariate

For parsimony we seek the coefficient function $\beta(t)$ in the following set of sparse step functions

$$\mathcal{E}_K = \left\{ \sum_{k=1}^{K} b_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}\left\{t \in \mathcal{I}_k\right\} : \ \mathcal{I}_1, \ldots, \mathcal{I}_K \text{ intervals } \subset \mathcal{T}, b_1, \ldots, b_K \in \mathbb{R} \right\} \tag{2.4}$$

where K is a hyperparameter that counts the number of intervals required to define the function. Note that we do not make any assumptions regarding the intervals $\mathcal{I}_1, \ldots, \mathcal{I}_K$. First they do not form a partition of $\mathcal{T}$. As a consequence, a function $\beta(t)$ in $\mathcal{E}_K$ is piecewise constant and null outside the union of the intervals $\mathcal{I}_k$, $k = 1, \ldots, K$. This union is the support of $\beta(t)$, hence the model includes an explicit description of the support. Second the intervals $\mathcal{I}_1, \ldots, \mathcal{I}_K$ can even overlap to ease the parametrization of the intervals: we do not have to add constraints on the parametrization to remove possible overlaps.

Now if we pick a function $\beta(t) \in \mathcal{E}_K$ with

$$\beta(t) = \sum_{k=1}^{K} b_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}\left\{t \in \mathcal{I}_k\right\}, \tag{2.5}$$

the integral of the covariate functions $x_i(t)$ against $\beta(t)$ becomes a linear combination of partial integrals of the covariate function over the intervals $\mathcal{I}_k$ and we predict y_i with

$$\widehat{y_i} = \mu + \sum_{k=1}^{K} b_k \, x_i(\mathcal{I}_k), \quad \text{where } x_i(\mathcal{I}_k) = \frac{1}{|\mathcal{I}_k|} \int_{\mathcal{I}_k} x_i(t) \mathrm{d}t.$$

Thus, given the intervals $\mathcal{I}_1, \ldots, \mathcal{I}_K$, we face a multivariate linear model with the usual Gaussian likelihood.

Then we set the parameters on $\mathcal{E}_K$ and a prior distribution. Each interval $\mathcal{I}_k$ is set with its center m_k and its half length ℓ_k:

$$\mathcal{I}_k = [m_k - \ell_k, m_k + \ell_k]. \tag{2.6}$$

As a result, when K is fixed, the parameter of the model is

$$\theta = (m_1, \ldots, m_K, \ell_1, \ldots, \ell_K, b_1, \ldots, b_K, \mu, \sigma^2).$$

Below, we denote $\beta_\theta(\cdot)$ the coefficient function defined with (2.5) to highlight the dependence on θ.

We first define the prior on the support, that is to say on the intervals $\mathcal{I}_k$. The prior of the center of each interval is uniformly distributed on the whole range of time $\mathcal{T}$. This uniform prior does not promote any particular region of $\mathcal{T}$. Furthermore, the prior of the half-length of the interval $\mathcal{I}_k$ is the Exponential distribution $\mathcal{E}(a)$. To understand this prior and set hyperparameters a, we introduce the prior probability that a given $t \in \mathcal{T}$ is in the support, namely

$$\alpha(t) = \int_{\Theta_K} \mathbf{1}\{t \in S_\theta\} \pi_K(\theta) \mathrm{d}\theta \tag{2.7}$$

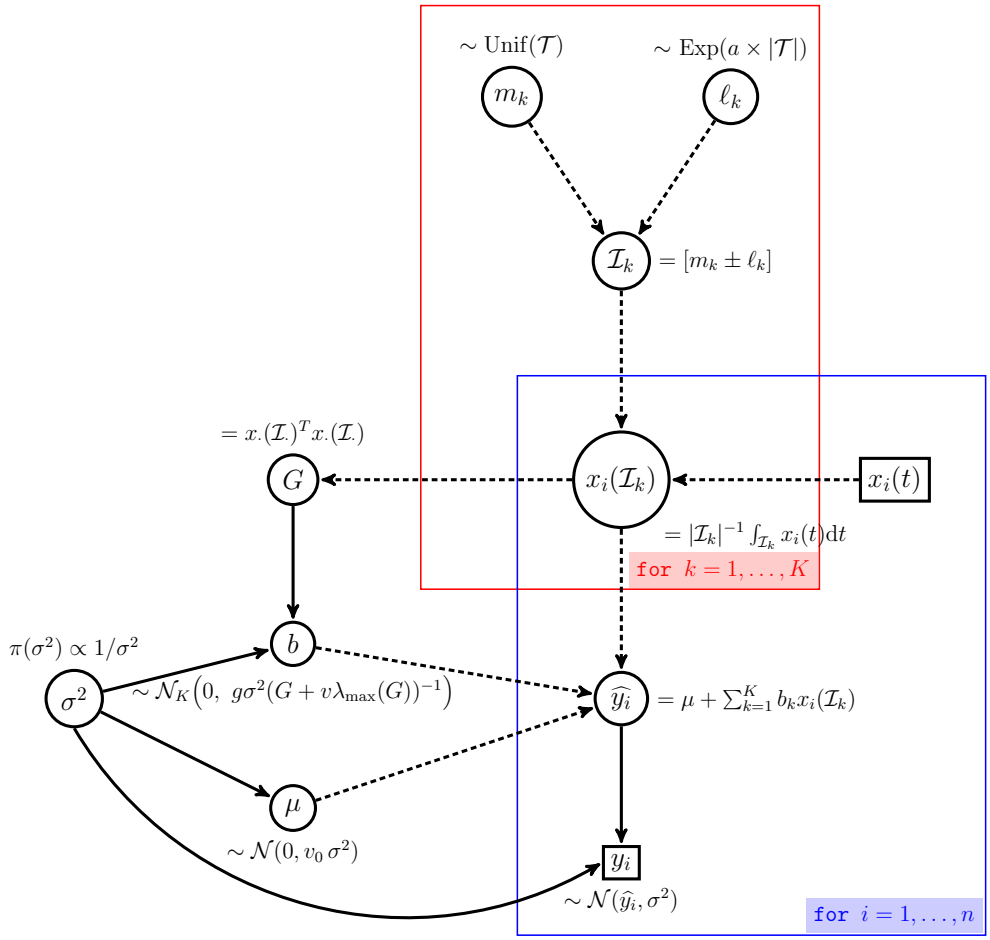


Figure 2.1: The full Bayesian model. *The coefficient function* $\beta(t) = \sum_{k=1}^K b_k \mathbf{1}\{t \in \mathcal{I}_k\}/|\mathcal{I}_k|$ *defines both a projection of the covariate functions* $x_i(t)$ *onto* $\mathbb{R}^K$ *by averaging the function over each interval* $\mathcal{I}_k$ *and a prediction* $\widehat{y}_i$ *which depends on the vector* $b = (b_1, \ldots, b_K)$ *and the intercept* μ*.*

where π_K is the prior distribution on the range of parameters Θ_K of dimension $3K + 2$, and where $S_\theta = \mathrm{Supp}(\beta_\theta)$ is the support of $\beta_\theta(t)$ that is to say the union of the $\mathcal{I}_k$. The value of $\alpha(t)$ depends on hyperparameters a. These parameters should be fixed with the help of prior knowledge on $\alpha(t)$.

Given the intervals, or equivalently, given the m_k and ℓ_k, the functional linear model becomes a multivariate linear model with $x_i(\mathcal{I}_k)$ as scalar covariates. We could have set a standard and well-understood prior on $b|(\mathcal{I}_k)_{1\leq k\leq K}$, namely the g-Zellner prior, with $g = n$ in order to define a vaguely informative prior. More specifically, the design matrix given the intervals is

$$x_\cdot(\mathcal{I}_\cdot) = \{x_i(\mathcal{I}_k),\ 1 \leq i \leq n,\ 1 \leq k \leq K\}\,.$$

And the g-Zellner prior, with $g = n$ is given by

$$\pi(\sigma^2) \propto 1/\sigma^2, \quad b|\sigma^2 \sim \mathcal{N}_K\left(0, n\sigma^2 G^{-1}\right) \tag{2.8}$$

where $b = (b_1, \ldots, b_K)$ and $G = x_\cdot(\mathcal{I}_\cdot)^T x_\cdot(\mathcal{I}_\cdot)$ is the Gram matrix. However, depending on the intervals $\mathcal{I}_k$, the covariates $x_i(\mathcal{I}_k)$ can be highly correlated. (We recall here that

the functional covariate can have autocorrelation and that the intervals can overlap.) That is why, in this setting, the Gram matrix $G = x_{\cdot}(\mathcal{I}_{\cdot})^T x_{\cdot}(\mathcal{I}_{\cdot})$ can be ill-conditioned, that is to say not numerically invertible and we cannot resort to the g-Zellner prior. To solve this problem we have to decrease the condition number of G, and apply a Tikhonov regularization. The resulting prior is a Ridge-Zellner prior (Baragatti and Pommeret, 2012) which replaces G by $G + \eta I$ in (2.8), where η is a scalar tuning the amount of regularization and I is the identity matrix. Adding the ηI matrix shifts all eigenvalues of the Gram matrix by η. In order to obtain a well-conditioned matrix, we decided to fix η with the help of the largest eigenvalue of the Gram matrix, $\lambda_{\max}(G)$ and to set $\eta = v\lambda_{\max}(G)$ where v is a hyperparameter of the model.

To sum up the above, the prior distribution on Θ_K is

$$\begin{aligned}
\mu|\sigma^2 &\sim \mathcal{N}\left(0, v_0\sigma^2\right), \\
b|\sigma^2, m_1, \ldots, m_K, \ell_1, \ldots, \ell_K &\sim \mathcal{N}_K\left(0, n\sigma^2(G + v\lambda_{\max}(G)I)^{-1}\right), \text{ where } G = x_{\cdot}(\mathcal{I}_{\cdot})^T x_{\cdot}(\mathcal{I}_{\cdot}), \\
\pi(\sigma^2) &\propto 1/\sigma^2, \\
m_k &\overset{i.i.d.}{\sim} \text{Unif}\left(\mathcal{T}\right), \quad k = 1, \ldots, K, \\
\ell_k &\overset{i.i.d.}{\sim} \text{Exp}(a \times |\mathcal{T}|), \quad k = 1, \ldots, K,
\end{aligned} \tag{2.9}$$

The resulting Bayesian modelling is given in Figure 2.1 and depends on hyperparameters which are v_0, v, a and K. We denote by $\pi_K(\theta)$ and $\pi_K(\theta|\mathcal{D})$ the prior and the posterior distributions. We propose below default values for the hyperparameters v_0, v, a. See Section 2.3.4 for numerical results that support this proposal.

- The parameter v_0 drives the prior information we put on the intercept μ. This is clearly not the most important hyperparameter since we expect important information regarding μ in the likelihood. We recommend using $v_0 = 100 \times \bar{y}^2$, where $\bar{y}$ is the average of the outcome on the dataset. Even if it may look like we set the prior with the current data, the resulting prior is vaguely non-informative.
- The parameter v is more difficult to set: it tunes the amount of regularization in the g-Zellner prior. Our set of numerical studies indicates, see Section 2.3 below, that $v = 5$ is a good value.
- The parameter a sets the prior length of an interval of the support up to a constant. This should depend on the number K of intervals. We recommend the value $a = 5K$ so that the average length of an interval from the prior distribution is proportional to $1/K$. Our numerical studies show that the choice of this constant 5 in the above recommendation does not drastically influence the results.

2.2.3 Model Choice

The hyperparameter K drives the number of intervals, thus the dimension of Θ_K. We can put an extra prior distribution on K and perform Bayesian model choice either to infer K or to aggregate posteriors coming from various values of K. There is a ban on the use of improper prior together with Bayesian model choice (or Bayes factor) because of

the Jeffrey-Lindley paradox (see, e.g. Robert, 2007, Section 5.2). A careful reader would notice here the improper prior on σ^2, but this does not prohibit the use of Bayesian choice because it is a parameter common to all models (i.e., to all values of K here).

Note that the marginal of the posterior distribution on a given interval (b_k, m_k, ℓ_k), $k \in \{1, \ldots, K\}$, is a multimodal distribution of dimension 3, with constraints on the support. Indeed, the intervals are exchangeable both a priori and a posteriori: we face a label switching issue. Moreover, the posterior distribution on the whole set of intervals is correlated: when (b_1, m_1, ℓ_1) is around one mode, the other intervals are around the other modes. Thus, the posterior distribution has a complex shape. Standard techniques such as harmonic mean or importance sample (Marin and Robert, 2010) that aim at computing the evidence of a model, namely $\pi(\mathcal{D}|K)$, or the Bayes factor, are difficult to carry out. This problem deserves another study. Regarding bridge sampling (Gelman and Meng, 1998), the main difficulty is that introducing a new interval in the model increases the dimension by 3. Running this efficient algorithm is thus not trivial at all in our context.

Nevertheless model information criteria such as AIC, BIC and DIC are much easier to compute. In this study, we have eliminated the Akaike Information Criterion (AIC) since it is designed to provide the model with the best predictive power. We have also eliminated the deviance information criterion (DIC) because this last criterion makes sense only when the posterior distribution is unimodal. (Our posterior distributions are much more complex, see above.) We thus recommend the use of the Schwartz information criterion (BIC) whose performance on our simulations was relatively good. But, as expected, when the size of the dataset is rather small or when the autocorrelation within the covariates is high, BIC tends to under-estimate the value of K (see supplementary materials).

2.2.4 Estimation of the Support

Regarding the inference of the support, an interesting quantity is the posterior probability that a given $t \in \mathcal{T}$ is in the support. It can be defined as the prior probability in (2.7), that is to say

$$\alpha(t|\mathcal{D}) = \int_{\Theta_K} \mathbf{1}\{t \in S_\theta\} \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta. \tag{2.10}$$

Both functions $\alpha(t)$ and $\alpha(t|\mathcal{D})$ can be easily computed with a sample from the prior and the posterior respectively. They are also relatively easy to interpret in terms of marginal distribution of the support: fix $t \in \mathcal{T}$,

- $\alpha(t)$ is the prior probability that t is in the support of the coefficient function and
- $\alpha(t|\mathcal{D})$ is the posterior probability of the same event.

Now let $L_\gamma(S, S_\theta)$ be the loss function given by

$$L_\gamma(S, S_\theta) = \gamma \int_0^1 \mathbf{1}\{t \in S \setminus S_\theta\} \mathrm{d}t + (1-\gamma) \int_0^1 \mathbf{1}\{t \in S_\theta \setminus S\} \mathrm{d}t \tag{2.11}$$

where $S_\theta = \mathrm{Supp}(\beta_\theta)$ is the support of $\beta_\theta(t)$, the coefficient function as parametrized in (2.5) and where γ is a tuning parameter in $[0, 1]$. Actually, there are two types of error when estimating the support:

- type I error: a point $t \in \mathcal{T}$ which is really in the support S_θ has not been included in the estimate,
- type II error: a point $t \in \mathcal{T}$ has been included in the support estimate but does not lie inside the real support S_θ

and the tuning parameter γ allows us to set different weights on both types of error. Note that, when $\gamma = 1/2$, the loss function is one half of the Lebesgue measure of the symmetric difference $S \Delta S_\theta$.

Bayes estimates are obtained by minimizing a loss function integrated with respect to the posterior distribution, see Robert (2007). Hence, in this situation, Bayes estimates of the support are given by

$$\widehat{S}_\gamma(\mathcal{D}) \in \arg\min_{S \subset \mathcal{T}} \int_{\Theta_K} L_\gamma(S, S_\theta) \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta. \tag{2.12}$$

The following theorem shows the existence of the Bayes estimate and how to compute it from $\alpha(t|\mathcal{D})$.

Théorème 2.1. *The level set of $\alpha(t|\mathcal{D})$ defined by*

$$\widehat{S}_\gamma(\mathcal{D}) = \{t \in \mathcal{T} \ : \ \alpha(t|\mathcal{D}) \geq \gamma\}$$

is a Bayes estimate associated with the above loss $L_\gamma(S, S_\theta)$. Moreover, up to a set of null Lebesgue measure, any Bayes estimate $\widehat{S}_\gamma(\mathcal{D})$ that solves the optimisation problem given in (2.12) *satisfies*

$$\{t \in \mathcal{T} \ : \ \alpha(t|\mathcal{D}) > \gamma\} \subset \widehat{S}_\gamma(\mathcal{D}) \subset \{t \in \mathcal{T} \ : \ \alpha(t|\mathcal{D}) \geq \gamma\}.$$

The proof of the above theorem is given in Appendix 2.6.1. Although simple-looking, the proof requires some caution because sets should be Borelian sets. Note that, when we try to completely avoid errors of type I (resp. type II) by setting $\gamma = 0$ (resp. $\gamma = 1$), the support estimate is $\mathcal{T}$ (resp. $\emptyset$). Additionally Theorem 2.1 shows how we should interpret the posterior probability $\alpha(t|\mathcal{D})$ and that its plot may be one important output of the Bayesian analysis proposed in this paper: it measures the evidence that a given point is in the support of the coefficient function.

Remark : Note that the number of intervals in the support estimate $\widehat{S}_\gamma(\mathcal{D})$ can be, and is often different from the value of K (because intervals can overlap). Therefore, the choice of the hyperparameter K (the number of intervals) can be validated with regard to the estimate $\widehat{S}_\gamma(\mathcal{D})$.

2.2.5 Estimation of the Coefficient Function

The Bayesian modelling given in Section 2.2.2 was mainly designed to estimate the support of the coefficient function. Bayes estimates of the coefficient function can be

made, and two alternatives are proposed below. The first one, given in Equation (2.13) is a smooth estimate, whereas the second estimate, given in Proposition 2.3, is a stepwise estimate which is parsimonious and may be more easily interpreted.

With the default quadratic loss, a Bayes estimate is defined as

$$\widehat{\beta}_{L^2}(\cdot) \in \underset{d(\cdot)\in L^2(\mathcal{T})}{\arg\min} \iint \left(\beta_\theta(t) - d(t)\right)^2 \mathrm{d}t \; \pi_K(\theta|\mathcal{D})\mathrm{d}\theta \tag{2.13}$$

where $\beta_\theta(t)$ is the coefficient function as parametrized in (2.5). At least heuristically $\widehat{\beta}_{L^2}(\cdot)$ is the average of $\beta_\theta(\cdot)$ over the posterior distribution $\pi_K(\theta|\mathcal{D})$, though the average of functions taking values in $L^2(\mathcal{T})$ under some probability distribution is hard to define (using either Bochner or Pettis integrals). In this simple setting we can claim the following (see Appendix 2.6.1 for the proof).

Proposition 2.2. *Let $\|\cdot\|$ be the norm of $L^2(\mathcal{T})$. If $\int \|\beta_\theta(\cdot)\| \, \pi_K(\theta|\mathcal{D})\mathrm{d}\theta < \infty$, then the estimate defined by*

$$\widehat{\beta}_{L^2}(t) = \int \beta_\theta(t)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta, \qquad t \in \mathcal{T}, \tag{2.14}$$

is in $L^2(\mathcal{T})$ and solves the optimization problem (2.13)*.*

Below, we call $\hat{\beta}_{L^2}$ the L^2-estimate. Averages such as (2.14) belong to the closure of the convex hull of the support $\mathcal{E}_K$ of the posterior distribution. We can prove (see Proposition 2.5 in Appendix 2.6.1) that the convex hull of $\mathcal{E}_K$ is the set $\mathcal{E} = \cup_{K=1}^{\infty}\mathcal{E}_K$ of step functions on $\mathcal{T}$, and the closure of $\mathcal{E}$ is $L^2(\mathcal{T})$. Hence the only guarantee we have on $\widehat{\beta}_{L^2}$ as defined in (2.14) is that $\widehat{\beta}_{L^2}$ lies in $L^2(\mathcal{T})$, a much larger space than the set of step functions. Though not shown here, integrating the $\beta_\theta(t)$'s over θ with respect to the posterior distribution has regularizing properties, and the Bayes estimate $\widehat{\beta}_{L^2}(t)$ is smooth.

To obtain an estimate lying in the set of step functions, namely $\mathcal{E}$, we can consider the projection of $\widehat{\beta}_{L^2}$ onto the set $\mathcal{E}_{K_0}$ for a suitable value of K_0 possibly different to K. However, due to the topological properties of $L^2(\mathcal{T})$ and $\mathcal{E}_{K_0}$, the projection of $\widehat{\beta}_{L^2}$ onto the set $\mathcal{E}_{K_0}$ does not always exist (see Appendix 2.6.1). To address this problem, we introduce a subset $\mathcal{E}^{\varepsilon}_{K_0}$ of $\mathcal{E}_{K_0}$, where $\varepsilon > 0$ is a tuning parameter. Let $\mathcal{F}^{\varepsilon}$ denote the set of step functions $\beta(t) \in L^2(\mathcal{T})$ which can be written as

$$\beta(t) = \sum b_k^{\dagger} \mathbf{1}\{t \in J_k\}$$

where the intervals J_k are mutually disjoint and each of the lengths are greater than ε. The set $\mathcal{E}^{\varepsilon}_{K_0}$ is now defined as $\mathcal{F}^{\varepsilon} \cap \mathcal{E}_{K_0}$. By considering this set, we remove from $\mathcal{E}_{K_0}$ the step functions which have intervals of very short length, and we can prove the following.

Proposition 2.3. *Let $K_0 \geq 1$ and $\varepsilon > 0$.*

(i) The function $d(\cdot) \mapsto \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|^2$ admits a minimum on $\mathcal{E}^{\varepsilon}_{K_0}$. Thus a projection of $\widehat{\beta}_{L^2}(\cdot)$ onto this set, defined by

$$\widehat{\beta}^{\varepsilon}_{K_0}(\cdot) \in \underset{d(\cdot)\in\mathcal{E}^{\varepsilon}_{K_0}}{\arg\min} \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|^2, \tag{2.15}$$

always exists.

(ii) The estimate $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ is a true Bayes estimate with loss function

$$L^{\varepsilon}_{K_0}\Big(d(\cdot),\beta(\cdot)\Big) = \begin{cases} \|d(\cdot)-\beta(\cdot)\|^2 = \int_{\mathcal{T}} \left(\beta(t)-d(t)\right)^2 \mathrm{d}t & \text{if } \beta \in \mathcal{E}^{\varepsilon}_{K_0}, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.16}$$

That is to say

$$\widehat{\beta}^{\varepsilon}_{K_0}(\cdot) \in \underset{d(\cdot)\in L^2(\mathcal{T})}{\arg\min} \int L^{\varepsilon}_{K_0}\Big(d(\cdot),\beta_\theta(\cdot)\Big)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta.$$

We call $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ the Bliss estimate given in Proposition 2.3. Finally one should note that the support of the Bliss estimate given in Proposition 2.3 provides another estimate of the support, which differs from the Bayes estimate introduced in Section 2.2.4. Obviously, real Bayes estimates, which optimize the loss integrated over the posterior distribution, are by construction better estimates. Another possible alternative would be the definition of an estimate of the coefficient function whose support is given by one of the Bayes estimates defined in Theorem 2.1. But such estimates do not account for the inferential error regarding the support. Hence we believed that, when it comes to estimating the coefficient function, the Bayes estimates proposed in this Section are better than other candidates and achieve a tradeoff between inferential errors on its support and prediction accuracy on new data.

2.2.6 Model with Several Functional Covariates

Suppose now we have not only observed a single functional covariate but q functional covariates $x_{ij}(t)$ defined on $\mathcal{T}$, for $i = 1,\dots,n$ and $j = 1\dots,q$. The model we have to study is

$$y_i|x_{i1}(t),\dots,x_{iq}(t),\mu,\beta_1(t),\dots,\beta_q(t),\sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + \sum_{j=1}^{q}\int \beta_j(t)x_{ij}(t)\mathrm{d}t\,,\ \sigma^2\right), \tag{2.17}$$

for $i = 1,\dots,n$ and $j = 1,\dots,q$. As in Section 2.2.2, each coefficient function $\beta_j(\cdot)$ is assumed to be a step function. In particular, for given $K_1,\dots,K_q$, we set $\beta_j(\cdot) \in \mathcal{E}_{K_j}$ for $j = 1,\dots,q$. Hence we have $\beta_j(t) = \sum_{k=1}^{K_j} b_{kj}\,\mathbf{1}\left\{t \in \mathcal{I}_{kj}\right\}/|\mathcal{I}_{kj}|$ where the $\mathcal{I}_{kj}$ are intervals of $\mathcal{T}$. Then, the outcome values y_i are predicted with

$$\hat{y}_i = \mu + \sum_{j=1}^{q}\sum_{k=1}^{K_j} b_{kj}x_{ij}(\mathcal{I}_{kj}), \qquad \text{where } \mathcal{I}_{kj} = \frac{1}{|\mathcal{I}_{kj}|}\int_{\mathcal{I}_{kj}} x_{ij}(t)\mathrm{d}t.$$

Hence, for given $K_1,\dots,K_q$, the parameter of the model is

$$\theta = (\theta_1,\dots,\theta_q,\mu,\sigma^2), \qquad \text{where } \theta_j = (m_{1j},\dots,m_{K_jj},\ell_{1j},\dots,\ell_{K_jj},b_{1j},\dots,b_{K_jj}).$$

Below, we denote by $\beta_{\theta,j}(\cdot)$ the j^{th} coefficient function defined with (2.5), which depends on θ. If we denote $K = \sum_{j=1}^{q} K_j$, the range of the parameter θ is denoted by Θ_K of which

the dimension is $3K+2$. The prior distribution on Θ_K is set in the same way as in Section 2.2.2:

$$\begin{aligned}
\mu|\sigma^2 &\sim \mathcal{N}\left(0, v_0\sigma^2\right), \\
b_j|\sigma^2, m_{1j}, \ldots, m_{K_jj}, \ell_{1j}, \ldots, \ell_{K_jj} &\sim \mathcal{N}_{K_j}\left(0, n\sigma^2(G_j + v\lambda_{\max}(G_j)I)^{-1}\right), \qquad j = 1, \ldots, q, \\
\pi(\sigma^2) &\propto 1/\sigma^2, \\
m_{kj} &\overset{i.i.d.}{\sim} \text{Unif}\left(\mathcal{T}\right), \quad k = 1, \ldots, K_j \text{ and } j = 1, \ldots, q, \\
\ell_{kj} &\overset{i.i.d.}{\sim} \text{Exp}(a \times |\mathcal{T}|), \quad k = 1, \ldots, K_j \text{ and } j = 1, \ldots, q,
\end{aligned} \tag{2.18}$$

where $b_j = (b_{1j}, \ldots, b_{K_jj})$ and G_j is given by $x_{\cdot j}(\mathcal{I}_{\cdot j})^T x_{\cdot j}(\mathcal{I}_{\cdot j})$ with

$$x_{\cdot j}(\mathcal{I}_{\cdot j}) = \{x_{ij}(\mathcal{I}_{kj}), 1 \leq i \leq n, 1 \leq k \leq K_j\}.$$

Below, we denote by $\pi_K(\theta)$ and $\pi_K(\theta|\mathcal{D})$ the prior and the posterior distributions. The estimators of the coefficient functions and their supports are defined as in Section 2.2.4 and 2.2.5 in the case of a single functional covariate. We denote by $S_{\theta,j}$ the support of $\beta_{\theta,j}(\cdot)$ which we estimate with

$$\widehat{S}_{\gamma,j}(\mathcal{D}) \in \underset{S \subset \mathcal{T}}{\arg\min} \int_{\Theta_K} L_\gamma(S, S_{\theta,j}) \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta,$$

where the loss function L_γ is given by (2.11) and for a fixed $\gamma \in (0,1)$.

The coefficient function $\beta_{\theta,j}(t)$ is estimated by using the estimators described in Proposition 2.2. The first one is:

$$\widehat{\beta}_{L^2,j}(\cdot) \in \underset{d(\cdot) \in L^2(\mathcal{T})}{\arg\min} \iint \left(\beta_{\theta,j}(t) - d(t)\right)^2 \mathrm{d}t \; \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta.$$

The second estimate is defined in the same vein by adapting the notation of Proposition 2.3:

$$\widehat{\beta}^{\varepsilon}_{K_0,j}(\cdot) \in \underset{d(\cdot) \in L^2(\mathcal{T})}{\arg\min} \int L^{\varepsilon}_{K_0}\Big(d(\cdot), \beta_{\theta,j}(\cdot)\Big) \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta.$$

2.2.7 Implementation

The full posterior distribution can be written explicitly from the Bayesian model given in Equations (2.9). As usual with hierarchical models, sampling from the posterior distribution $\pi_K(\theta|\mathcal{D})$ can be done with a Gibbs algorithm (see, e.g., Robert and Casella, 2013, Chapter 7). The details of the MCMC algorithm are given in Appendix 2.6.2 for the case of one single functional covariate and for the case of several functional covariates.

Now, for simplicity of notation, we focus on the single functional covariate case. Let $\theta(s)$, $s = 1, \ldots, N$, denote the output of the MCMC sampler after the burn-in period. The computation of the Bayes estimate $\widehat{S}_\gamma(\mathcal{D})$ of the support as defined in Theorem 2.1 depends on the probabilities $\alpha(t|\mathcal{D})$. With the Monte Carlo sample from the MCMC, we can easily approximate these posterior probabilities by the frequencies

$$\alpha(t|\mathcal{D}) \approx \frac{1}{N} \sum_{s=1}^{N} \mathbf{1}\{\beta_{\theta(s)}(t) \neq 0\}.$$

What remains to be computed are the approximations of $\widehat{\beta}_{L^2}(\cdot)$ and $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ based on the MCMC sample. First, the Monte Carlo approximation of (2.14) is given by

$$\widehat{\beta}_{L^2}(t) \approx \frac{1}{N}\sum_{s=1}^{N} \beta_{\theta(s)}(t).$$

More interestingly, the Bayes estimate $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ can be computed by minimizing

$$\left\| d(\cdot) - \widehat{\beta}_{L^2}(t) \right\|^2$$

over the set $\mathcal{E}^{\varepsilon}_{K_0}$. To this end we run a Simulated Annealing algorithm (Kirkpatrick et al., 1983), described in Appendix 2.6.2.

We also provide a striking graphical display of the posterior distribution on the set $\mathcal{E}_K$ with a heat map. More precisely, the aim is to sketch all marginal posterior distributions $\pi^t_K(\cdot|\mathcal{D})$ of $\beta_\theta(t)$ for any value of $t \in \mathcal{T}$ in one single figure. To this end we introduce the probability measure Q on $\mathcal{T} \times \mathbb{R}$ defined as follows. Its marginal distribution over $\mathcal{T}$ is uniform, and given the value t of the first coordinate, the second coordinate is distributed according to the posterior distribution of $\beta_\theta(t)$. In other words,

$$(t,h) \sim Q \quad \Longleftrightarrow \quad t \sim \text{Unif}(\mathcal{T}),\ h|t \sim \pi^t_K(\cdot|\mathcal{D}).$$

We can easily derive an empirical approximation of Q from the MCMC sample $\{\theta(s)\}$ of the posterior. Indeed, the first marginal distribution of Q, namely Unif($\mathcal{T}$) can be approximated by a regular grid t_i, $i = 1, \ldots, M$. And, for each value of i, set $h_{is} = \beta_{\theta(s)}(t_i)$, $s = 1, \ldots, N$. The resulting empirical measure is

$$\widehat{Q} = \frac{1}{M\,N} \sum_{i=1,\ldots,M} \sum_{s=1,\ldots,N} \delta_{(t_i,h_{is})},$$

where $\delta_{(t,h)}$ is the Dirac measure at (t,h). The graphical display we propose is representing $\widehat{Q}$ with a heat map on $\mathcal{T} \times \mathbb{R}$. Each small area of $\mathcal{T} \times \mathbb{R}$ is thus coloured according to its $\widehat{Q}$-probability. This should be done cautiously as the marginal posterior distribution $\pi^t_K(\cdot|\mathcal{D})$ has a point mass at zero: $\pi^t_K(h = 0|\mathcal{D}) > 0$ by construction of the prior distribution. Finally the colour scale can be any monotone function of the probabilities, in particular non linear functions to handle the atom at 0. Examples are provided in Section 2.3 in Figures 2.4 and 2.5.

Remark In practice the whole function x_i may be unknown and only observed at a finite set of time points $\{t_{ij}, j = 1, \ldots, n_i\}$. The time points may be irregularly spaced and vary between individuals. This common situation of applied functional data analysis is usually handled by converting the discrete measures $\{x_i(t_{ij}), j = 1, \ldots, n_i\}$ to a function computable for any time point by using interpolation or smoothing techniques (see Staniswalis and Lee, 1998; Di et al., 2009 or Ramsay and Silverman, 2005 page 9 and chapter 15 of the second edition, or Crambes et al., 2009 page 41). In the present paper, it is worth noting that the whole curve x_i is actually not needed. The only requirement is to compute the value of the integral $\int_{\mathcal{I}_k} x_i(t)\mathrm{d}t$ for any given interval $\mathcal{I}_k$. Several numerical techniques are available for this purpose when the observed time points are irregular (see, for example, Deheuvels, 1980, Chapter V or Phythian and Williams, 1986; Yao et al., 2003). For simplicity, we choose the trapezoidal rule in the simulations as the derived precision is sufficient in our context.

2.3 Simulation Study

In this section, the performance of univariate Bliss is evaluated and compared to three competitors: BFDA (Crainiceanu and Goldsmith, 2010), Fused lasso (Tibshirani et al., 2005) and Flirti (James et al., 2009), using simulated datasets.

- The BFDA method aims to fit a Bayesian penalized B-splines model. The BFDA estimate minimizes the posterior expected L^2-loss, computed by using a Monte Carlo approximation from an MCMC sample. Moreover, in order to compare this Bayesian approach to Bliss, we compute a representation of the marginal posterior distributions (see Section 2.2.7) from the BFDA's MCMC sample.
- Fused Lasso is an approach based on minimizing a penalized likelihood in order to induce parsimony on the values $\beta(t)$ and on the differences $\beta(t) - \beta(t')$ when t and t' are close.
- Flirti proceeds in the same vein by introducing a penalization term which promotes parsimony on the coefficient function and its derivatives. Below, we apply Flirti by using a penalization term in such a way that its first derivate is sparsely estimated. Hence, the Flirti estimate should theoretically be a step function as the stepwise-Bliss estimate. Moreover, the authors propose to compute confidence bands by using a bootstrap procedure.

Below, Section 2.3.1 describes how we generate data sets with one single functional covariate. Then, the performances of the support estimate of the Bliss method are described in Section 2.3.2. We compare in Section 2.3.3 the coefficient function estimators of the different methods. In Section 2.3.4, we evaluate the sensitivity of the estimates with respect to the model's hyperparameters. Next the multivariate Bliss model defined in Section 2.2.6 is applied twice on simulated datasets with two uncorrelated functional covariates and then with two correlated functional covariates. We discuss the computational time of the algorithms described in Section 2.2.7 applied on the following simulated data sets in Appendix 2.6.3.

2.3.1 Simulation Scheme for Datasets with One Functional Covariate

First of all, we describe how we generate different datasets on which we applied and compared the methods. The support of the covariate curves $x_i(\cdot)$ is $\mathcal{T} = [0, 1]$ observed on a regular grid $\boldsymbol{t} = (t_1, \ldots, t_p)$ on $\mathcal{T}$, for $p = 100$. We simulate p-multivariate Gaussian vectors z_i, $i = 1, \ldots, n$ (with $n = 100$), corresponding to the values $x_i(t)$ for the observation times $t \in \boldsymbol{t}$. The covariance matrix Σ of these Gaussian vectors is derived from the covariance between $x_i(t)$ and $x_i(t')$ given by:

$$\sqrt{\mathrm{var}(t)\,\mathrm{var}(t')}\exp\left(-\zeta^2(t-t')^2\right),$$

where $var(t)$ is the variance of the values $x_i(t)$ for $i = 1, \ldots, n$ and the coefficient ζ tunes the autocorrelation of the curves $x_i(\cdot)$. Three different shapes are considered for the functional coefficient β, given in Figure 2.2.

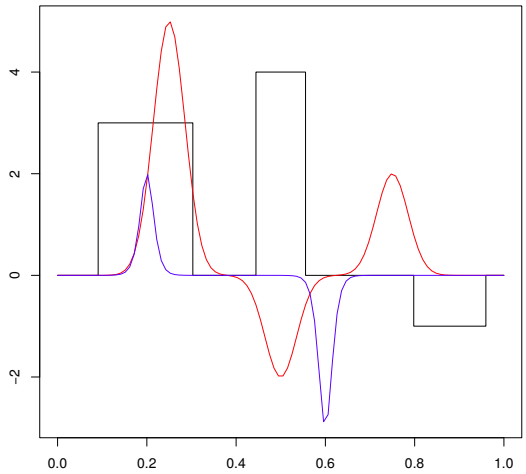
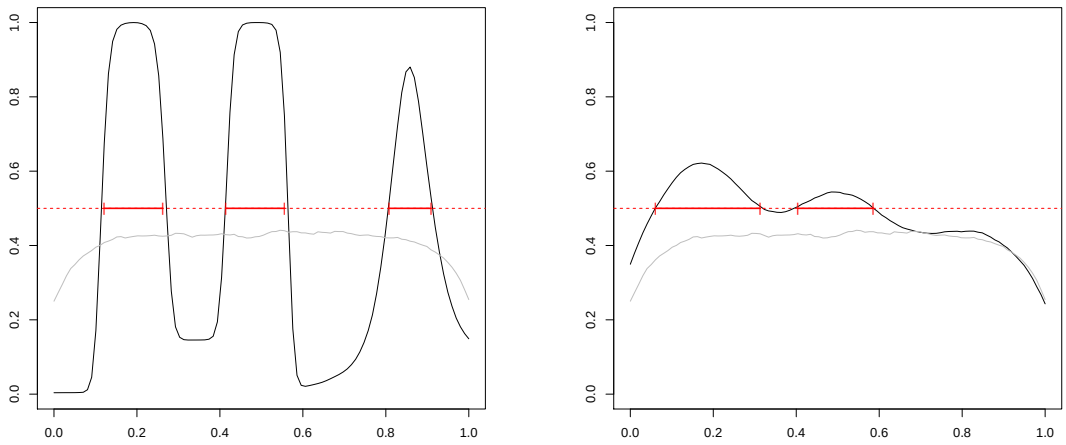


Figure 2.2: Coefficient functions for numerical illustrations. *The black (resp. red and blue) curve corresponds to the coefficient function of Shape "Step function" (resp. "Smooth" and "Spiky").*



Dataset 1 ($r = 5$, $\zeta = 1$) Dataset 3 ($r = 5$, $\zeta = 1/5$)

Figure 2.3: Prior (in gray) and posterior (in black) probabilities of being in the support computed on Datasets 1 and 2. *Bayes estimates of support using Theorem 2.1 with $\gamma = 1/2$ are given in red.*

The first one is a step function, the second one is smooth and is null on small intervals of $\mathcal{T}$ (Smooth), the third one is non-null only on small intervals of $\mathcal{T}$ (Spiky).

- Step function: $\beta(t) = 3\times\mathbf{1}\{t \in [0.1, 0.3]\}+4\times\mathbf{1}\{t \in [0.45, 0.55]\}-\mathbf{1}\{t \in [0.8, 0.95]\}$.
- Smooth: $\beta(t) = 5 \times e^{-20(t-0.25)^2} - 2 \times e^{-20(t-0.5)^2} + 2 \times e^{-20(t-0.75)^2}$.
- Spiky: $\beta(t) = 8 \times \left(2 + e^{20-100t} + e^{100t-20}\right)^{-1} - 12 \times \left(2 + e^{60-100t} + e^{100t-60}\right)^{-1}$.

The outcomes y_i are calculated according to (2.3). The value of σ^2 is fixed in such a way that the signal to noise ratio is equal to a chosen value r. Datasets are simulated for $\mu = 1$ and for the following different values of ζ and r:

- $\zeta = 1, 1/3, 1/5,$
- $r = 1, 3, 5.$

Hence, we simulate 27 datasets with different characteristics, that we use in Section 2.3.3 to compare the methods.

2.3.2 Performances Regarding Support Estimates

Table 2.1: *Comparison of the support estimate and the support of the Bliss estimate.*

			Support Error		Dataset
Shape	r	ζ	Support of the stepwise Bliss estimate	Bayes support estimate	
Step function	5	1	0.242	0.152	1
	5	1/3	0.384	0.202	2
	5	1/5	0.242	0.293	3
	3	1	0.232	0.091	4
	3	1/3	0.323	0.394	5
	3	1/5	0.424	0.465	6
	1	1	0.283	0.162	7
	1	1/3	0.404	0.333	8
	1	1/5	0.439	0.394	9

Section 2.3.1 describes the simulation scheme of the datasets and Section 2.3.3 describes the criteria: Support Error.

We begin by assessing the performances of our proposal in terms of support recovery. We focus here on the datasets simulated with the step function as the true coefficient function. It is the only function among the three functions we have chosen where the real definition of the support matches with the answer a statistician would expect, see Figure 2.2. The numerical results are given in Table 2.1, where we evaluated the error with the Lebesgue measure of the symmetric difference between the true support S_0 and the estimated one $\widehat{S}$, that is to say $2L_{1/2}(\widehat{S}, S_0)$ with the notation of Section 2.2.4.

As we claim at the end of Section 2.2.5, the Bayes estimate defined in Theorem 2.1 performs much better than the support of the Bliss estimate of Proposition 2.3. As also expected the accuracy of the Bayes support estimate worsens when the autocorrelation within the functional covariate $x_i(t)$ increases. The signal to noise ratio is the second most influent factor that explains the accuracy of the estimate.

The third interval of the true support, namely $[0.8, 0.95]$, is the most difficult to recover because the true value of the coefficient function over this interval is relatively low (-1) compared to the other values (4 and 3) of the coefficient function. Figure 2.3 gives two examples of the posterior probability function $\alpha(t|\mathcal{D})$ defined in Eq. (2.10) where we have highlighted (in red) the Bayes support estimate with $\gamma = 1/2$. Of these two examples, Figure 2.3 shows that the third interval is recovered only when there is low autocorrelation in the curves $x_i(\cdot)$ (i.e. Dataset 1). Figure 2.3 shows that the support estimate of Dataset 1 (low autocorrelation within the covariate) is more trustworthy than the support estimate of Dataset 3 (high autocorrelation within the covariate).

For more complex coefficient functions, see Figure 2.2, we cannot compare the Bayes support estimate directly with the true support of the coefficient function that generated

the data. Nevertheless, in the next section, we will compare the coefficient estimate with the true value of the coefficient function.

2.3.3 Performances Regarding the Coefficient Function

In order to compare the methods for the estimation of the coefficient function, we use the L^2-loss, namely

$$\int_0^1 (\widehat{\beta}(t) - \beta_0(t))^2 \mathrm{d}t \tag{2.19}$$

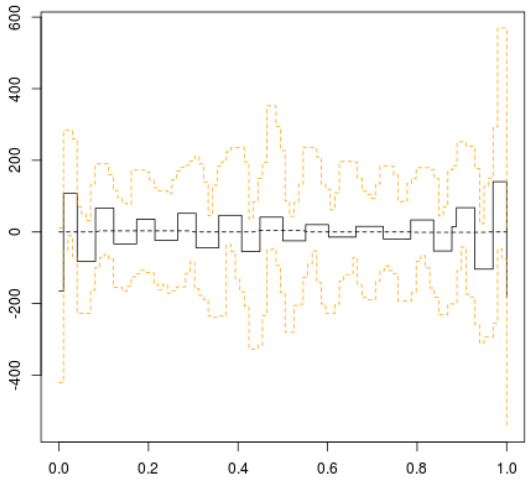
where $\widehat{\beta}(t)$ is an estimate we compare to the true coefficient function $\beta_0(t)$. Table 2.2 shows the results of Bliss and its competitors on these simulated datasets. It appears that the numerical results of the three methods have the same order of magnitude although the three methods may have different accuracy, depending on the shape of the coefficient function that generated the dataset.

Regarding Fused Lasso, we can see in Table 2.2 that its accuracy worsens when the problem is not sparse, that is to say when the true function is the "smooth" function (the red curve of Figure 2.2). Next, we observe that Flirti is very sensitive. Its numerical results can sometimes be quite accurate, but sometimes the L^2-error can blow up (to exceed 100) because the method did not manage to tune its parameters. Concerning the BFDA method, we note that the estimate becomes irrelevant when the autocorrelation increases (i.e. ζ decreases). In particular, for the Step function and Smooth shapes, the L^2-errors can exceed 10^3. The L^2-estimate defined in Proposition 2.2 frequently overperforms the other methods. This first conclusion is not surprising because the L^2-estimate has been defined to optimize the L^2-loss integrated over the posterior distribution.

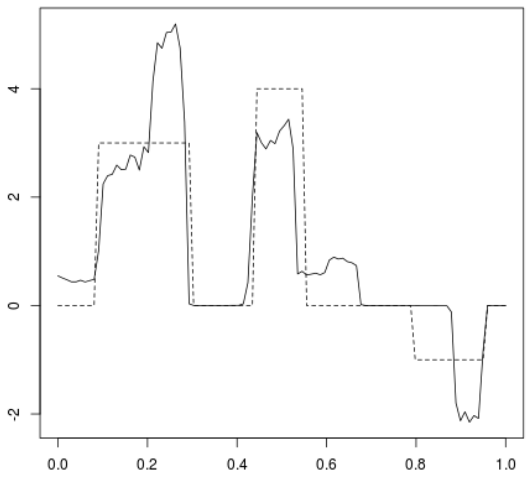
Even in situations where the true function is stepwise, the stepwise Bliss estimate of Proposition 2.3 is less accurate than the L^2-estimate, except for two examples (datasets 6 and 9). Nevertheless we do insist that the stepwise Bliss estimate was built to provide a trade off between accuracy regarding the support estimate and accuracy regarding the coefficient function estimate. Thus the stepwise estimate is a balance between support estimate and coefficient function estimate that can help the statistician who can then obtain an interpretation of the underlying phenomena that generated the data. In other words, the stepwise Bliss estimate is not the best either at estimating the support or at approximating the coefficient function, but provides a tradeoff.

To show more detailed results we have presented the estimate of the coefficient function in three cases.

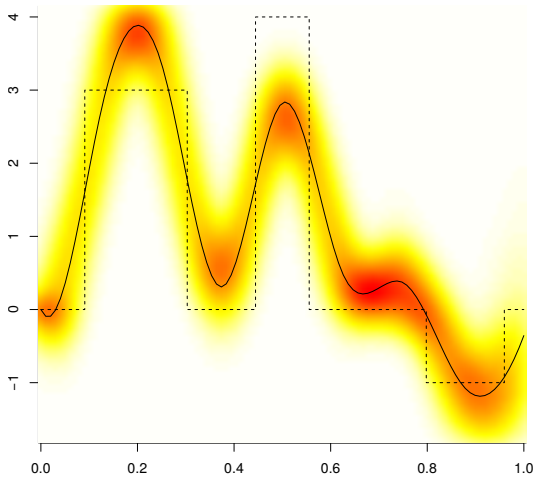
- Figure 2.4 displays the numerical results on Dataset 4 (medium level of signal, low level of autocorrelation within the covariate). As can be expected when the true coefficient is a stepwise function, the stepwise Bliss estimate behaves nicely. The representation of the marginals of the posterior distribution with a heat map shows the confidence we can have in the Bayes estimate of the coefficient function. The smooth L^2-estimate nicely follows the regions of high posterior density. Here, the stepwise estimate clearly highlights two time periods (the first two intervals of the true support) and the sign of the coefficient function on these intervals. We



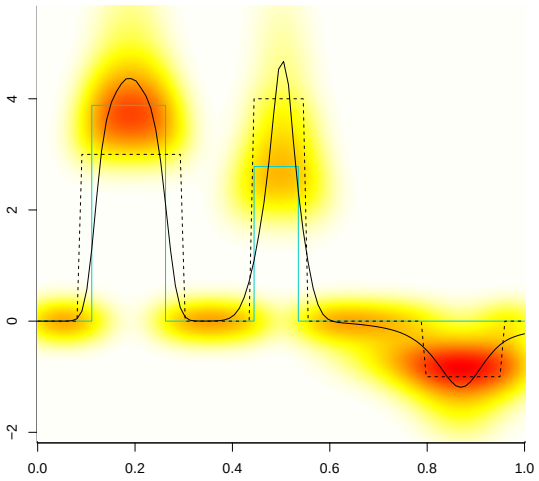
Flirti



Fused Lasso



BFDA



Bliss

Figure 2.4: Estimates of the coefficient function on Dataset 4 ($r = 3$, $\zeta = 1$). *For each plot, the black dotted line is the true coefficient function (Step function, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of $\beta(t)$ are represented using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate.*

can compare our proposal with its competitors. Flirti did not manage to tune its own parameters, and the Flirti estimate is irrelevant. Fused Lasso on a discretized version of the functional covariate provides a relatively nice estimate of the coefficient function. BFDA is not that bad, although the estimate is clearly too smooth to match the true coefficient function. As for the Bliss method, we give a representation

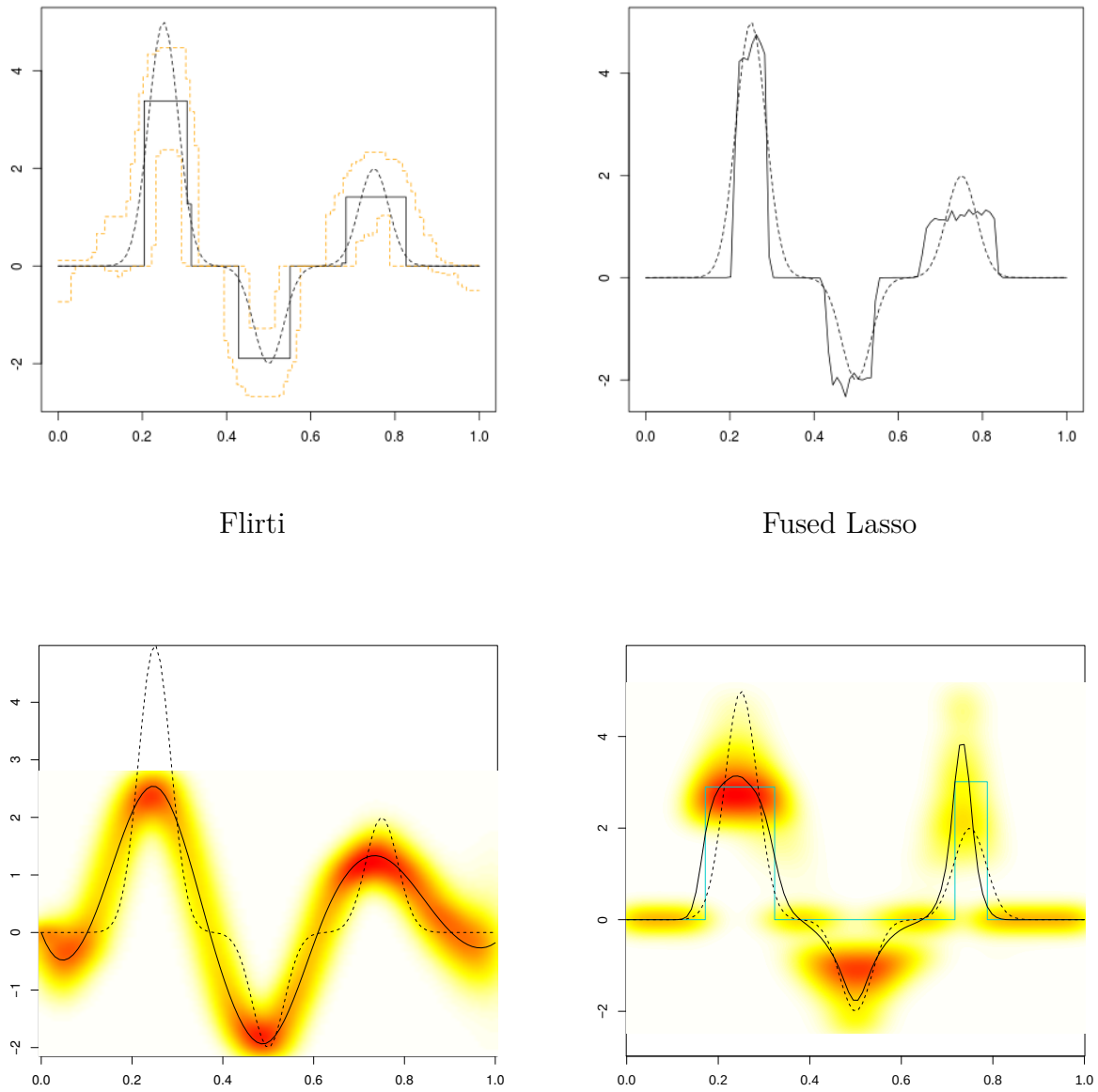


Figure 2.5: Estimates of the coefficient function on Dataset 13 ($r = 3$, $\zeta = 1$)). *For each plot, the black dotted line is the true coefficient function (Smooth, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of $\beta(t)$ are represented using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate.*

of the marginals of the posterior distribution based on the MCMC sample provided by the BFDA method. In this case, the true coefficient function is not in high posterior density regions, especially for the values of t for which it equals 0.

- Figure 2.5 displays the numerical results on Dataset 13 (medium level of signal, low

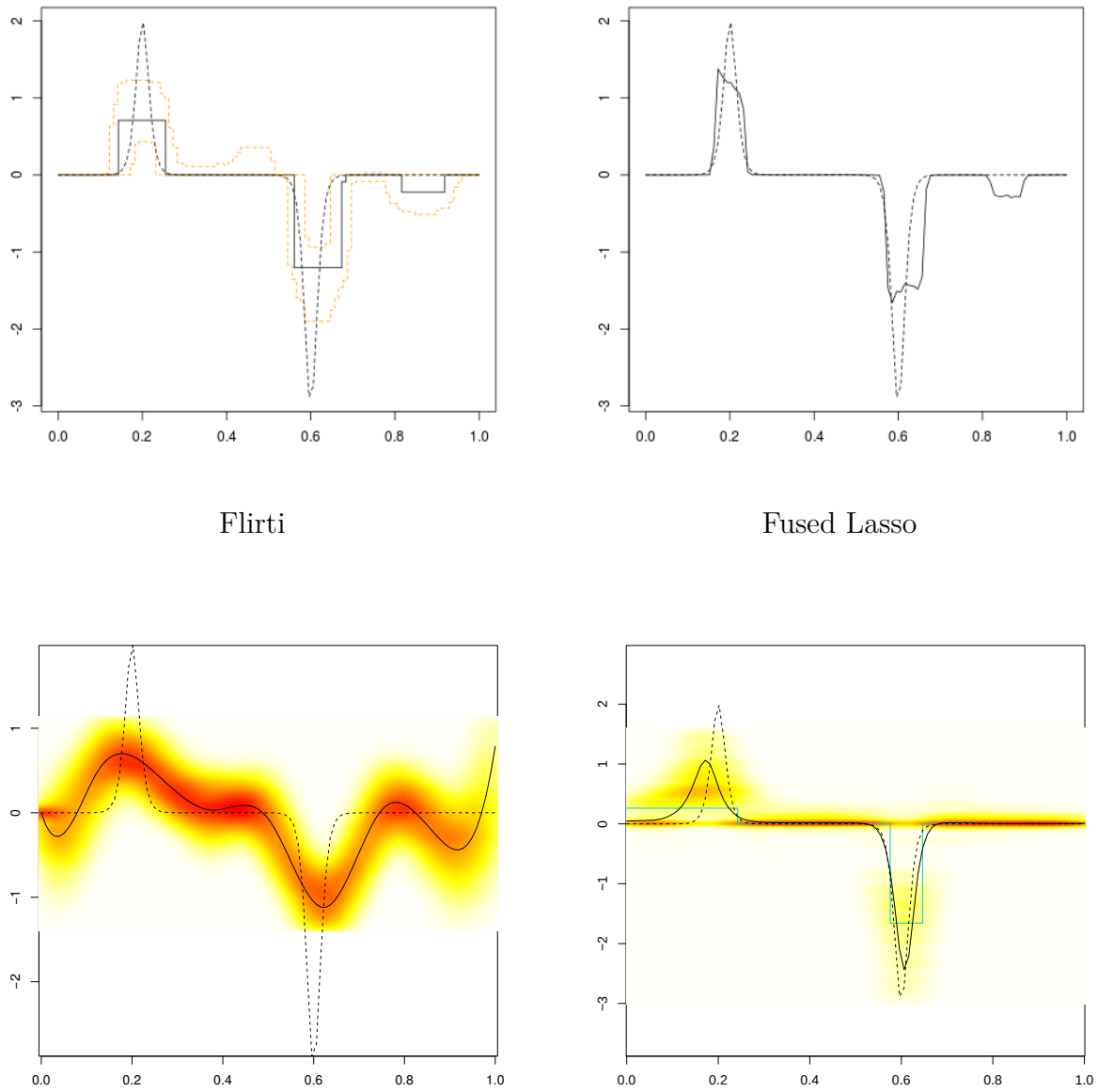


Figure 2.6: Estimates of the coefficient function on Dataset 25 ($r = 1$, $\zeta = 1$)). *For each plot, the black dotted line is the true coefficient function (Spiky, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of $\beta(t)$ are represented using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate.*

level of autocorrelation within the covariate). In this example, the true coefficient is not stepwise, but smooth, and is around zero on small time periods. Flirti and Fused Lasso performed nicely. Their estimates are exactly 0 on intervals where the true coefficient function is 0. Hence they clearly highlight the intervals. The BFDA estimate decently fits the true coefficient function but it is less accurate when the

Table 2.2: *Numerical results of Bliss, Flirti, Fused lasso and FDA on the Simulated Datasets.*

			L^2-error					Dataset
Shape	r	ζ	Bliss estimate	L^2-estimate	Fused lasso	Flirti	BFDA	
Step function	5	1	1.126	0.740	**0.666**	1.288	0.672	1
	5	1/3	2.221	**1.415**	1.947	1.781	10^5	2
	5	1/5	2.585	**1.656**	1.777	3.848	10^5	3
	3	1	1.283	0.821	0.984	10^3	**0.752**	4
	3	1/3	1.531	**1.331**	1.936	10^4	10^6	5
	3	1/5	2.266	2.989	2.036	**1.772**	10^5	6
	1	1	1.589	**0.747**	0.995	3.848	0.877	7
	1	1/3	2.229	**1.817**	2.214	10^4	10^6	8
	1	1/5	**1.945**	2.364	2.028	3.848	10^4	9
Smooth	5	1	0.510	**0.134**	0.601	0.166	0.553	10
	5	1/3	0.807	0.609	**0.442**	2.068	5.283	11
	5	1/5	1.484	**1.352**	2.325	2.068	10^4	12
	3	1	0.776	0.416	0.320	**0.263**	0.512	13
	3	1/3	**0.855**	0.954	6.790	2.068	4.782	14
	3	1/5	1.291	**1.162**	1.742	1.328	10^3	15
	1	1	0.932	0.641	0.652	2.335	**0.577**	16
	1	1/3	0.719	**0.283**	0.613	10^4	10^3	17
	1	1/5	1.536	**1.006**	4.680	5.430	10^3	18
Spiky	5	1	0.099	**0.013**	0.059	0.035	0.213	19
	5	1/3	0.208	**0.144**	0.260	0.271	0.501	20
	5	1/5	0.285	0.251	**0.181**	0.226	1.882	21
	3	1	0.187	**0.023**	0.638	0.136	0.207	22
	3	1/3	0.257	0.202	**0.159**	0.277	0.473	23
	3	1/5	0.269	**0.260**	0.459	0.276	5.416	24
	1	1	0.144	**0.087**	0.123	0.166	0.217	25
	1	1/3	0.242	**0.223**	0.260	10^2	0.675	26
	1	1/5	0.273	0.279	**0.221**	0.301	3.208	27

Section 2.3.1 describes the simulation scheme of the datasets. The stepwise Bliss estimate is the estimate defined in Proposition 2.3, while the L^2-estimate is the smooth estimate defined in Proposition 2.2.

coefficient function is around 0. The L^2-estimate performed as well as the BFDA estimate, except it is around 0 for t in $[0, 0.1]$ and $[0.9, 1]$. The stepwise Bliss estimate performed relatively poorly because it does not detect a negative interval around $t = 0.5$. With respect to Proposition 2.3, it is a projection of the L^2-estimate (which is clearly different to 0 around $t = 0.5$), hence we could expect that the stepwise Bliss estimate is negative for t around 0.5. Therefore, in this case, the simulated annealing algorithm does not correctly converge.

- Figure 2.6 displays the numerical results on Dataset 25 (low level of signal, and low level of autocorrelation within the covariate). In this example, the true coefficient is not stepwise, but smooth, and is around zero on large time periods. The L^2-estimate of Proposition 2.2 matches approximately the true coefficient function. The stepwise Bliss estimate is a little bit poorer (maybe because of the difficult calibration of the simulated annealing algorithm). When comparing these results with other estimates on this dataset, we see that Flirti and Fused Lasso performed decently also, even if they both highlight a third time period (around $t = 0.85$) where they infer a negative coefficient function instead of 0. In this case, Flirti has managed to tune its own parameters in a relevant way. The confidence bands of Flirti are therefore reliable, but we stress here that they are relatively wide around periods where the Flirti estimate is null and do not reflect high confidence in any support estimate based on Flirti. Finally, the comments on BFDA are the same as Dataset 4, the BFDA estimate has clearly been too smoothed to match the true coefficient function, especially for t for which it is around 0.

2.3.4 Tuning the Hyperparameters

We can now discuss our recommendation on the hyperparameters of the model, given at the end of Section 2.2.2. For this study, we applied our methodology on Dataset 1 and fixed the hyperparameters v_0, v, a around the recommended values. Remember that Dataset 1 is a synthetic dataset simulated with a coefficient function that is a step function (the black curve of Figure 2.2), with a high level of signal over noise ($r = 5$) and with a low level of autocorrelation within the covariate ($\zeta = 1$). The following values are considered for each hyperparameter:

- for a: $2K$, $5K$, $10K$, $15K$ and $20K$;
- for v: 10, 5, 2, 1 and 0.5;
- and for K: any integer between 1 and 10.

The numerical results are given in Table 2.3. The default values we recommend are not the best values here, but we have done numerous other trials on many synthetic datasets and these choices are relatively robust.

Table 2.3: *Performances of Bliss with respect to the tuning of the hyperparameters.*

	Error on the β		Error on the support	
	Bliss estimate	L^2-estimate	Support of the stepwise Bliss estimate	Bayes support estimate
$a = 2K$	1.000	0.698	0.222	0.439
$a = 5K$ ♡	1.013	1.135	0.222	0.192
$a = 10K$	1.642	1.364	0.242	0.202
$a = 15K$	3.060	1.645	0.364	0.212
$a = 20K$	2.032	1.888	0.263	0.263
$v = 10$	1.628	1.125	0.242	0.192
$v = 5$ ♡	1.711	1.131	0.242	0.192
$v = 2$	1.082	1.143	0.273	0.192
$v = 1$	1.207	1.119	0.273	0.192
$v = 0.5$	1.675	1.129	0.263	0.192
$K = 1$	1.798	1.782	0.424	0.449
$K = 2$	0.993	1.101	0.222	0.222
$K = 3$	1.696	1.124	0.242	0.192
$K = 4$	1.736	1.159	0.283	0.172
$K = 5$	2.081	1.233	0.303	0.172
$K = 6$	2.177	1.243	0.283	0.202
$K = 7$	2.135	1.221	0.303	0.232
$K = 8$	1.343	1.184	0.263	0.242
$K = 9$	1.439	1.166	0.263	0.328
$K = 10$	1.897	1.089	0.364	0.348

The ♡ symbol indicates the default values.

2.3.5 Simulation Study for Two Functional Covariates

Simulation scheme for datasets with two functional covariates

We describe how we generate datasets with two functional covariates. The curves $x_{i1}(\cdot)$ are generated on a regular grid $\boldsymbol{t^1} = (t^1_1, \ldots, t^1_{p_1})$ on $\mathcal{T}$, for $p_1 = 50$ and the curves $x_{i2}(\cdot)$

are generated on a regular grid $\boldsymbol{t^2} = (t_1^2, \ldots, t_{p_2}^2)$ on $\mathcal{T}$, for $p_2 = 100$. We simulate z_i a $(p_1 + p_2)$-multivariate Gaussian vectors for $i = 1, \ldots, n$ (with $n = 200$). The first p_1 coordinates of z_i define the values of the curve $x_{i1}(\cdot)$ for the observation times in $\boldsymbol{t^1}$. The last p_2 coordinates define the values of the curve $x_{2i}(\cdot)$ for the observation times in $\boldsymbol{t^2}$. Hence, $z_i = \Big(x_{i1}(t_1^1), \ldots, x_{i1}(t_{p_1}^1), x_{i2}(t_1^2), \ldots, x_{i2}(t_{p_2}^2)\Big)$ for each $i = 1, \ldots, n$. The covariance matrix Σ of the entire Gaussian vectors $z = \Big(z_1, \ldots, z_n\Big)$ is defined so that

1. for $j = 1, 2$, for t and t' in $\boldsymbol{t^j}$, the covariance between $x_{ij}(t)$ and $x_{ij}(t')$ is
$$\sqrt{\mathrm{var}_j(t)\,\mathrm{var}_j(t')}\exp\left(-\zeta^2(t-t')^2\right), \tag{2.20}$$
2. for $t \in \boldsymbol{t^1}$ and $t' \in \boldsymbol{t^2}$, the covariance between $x_{i1}(t)$ and $x_{i2}(t')$ is
$$c \times \sqrt{\mathrm{var}_1(t)\,\mathrm{var}_2(t')}\exp\left(-\zeta^2(t-t')^2\right), \tag{2.21}$$
for a given $c \in [-1, 1]$,

where $\mathrm{var}_j(t)$ is the variance of the $(x_{ij}(t))_{i=1,\ldots,n}$. The tuning parameter ζ in (2.20) drives the autocorrelation of curves $x_{ij}(\cdot)$ and below ζ is fixed to be 1. The tuning parameter c in (2.21) drives the cross-covariance between the curves $x_{i1}(\cdot)$ and $x_{i2}(\cdot)$. For $c = 0$, the curves $x_{i1}(\cdot)$ and $x_{i2}(\cdot)$ are uncorrelated and for c close to 1 the curves are highly correlated. Figure 2.8 shows examples of matrix Σ for $\zeta = 1$ and for different values of c.

The outcome values y_i are calculated according to (2.2) where $\beta_1(\cdot)$ and $\beta_2(\cdot)$ are the coefficient functions shown in Figure 2.7, and σ^2 is fixed so that the signal to noise ratio r is equal to 5. Four data sets are generated for $c = 0, 0.3, 0.6$ and 0.9 in order to illustrate how the estimates behave when the correlation between the functional covariates increases.

Below, we apply the model described in Section 2.2.6 with the default values for the hyperparameters: $K_1 = K_2 = 3$, $a = 5K$ and $v = 5$, as prescribed in Section 2.3.4. The results are given with Figures 2.9 and 2.10.

Performances regarding support estimates

Figure 2.9 shows the support estimates of $\beta_1(\cdot)$ and $\beta_2(\cdot)$ for uncorrelated covariates ($c = 0$) and for highly correlated covariates ($c = 0.9$). For $c = 0$ (Plots (a) and (b)), we notice that the support estimates approximately find the two positive intervals but do not find the third interval, for the first covariate as for the second one. For $c = 0.9$ (Plots (c) and (d)), the $\beta_2(\cdot)$ support estimate fails to detect the second one. We suspect that this is due to the high correlation between the two covariates.

Performances regarding the coefficient function

Figure 2.10 shows the estimators of $\beta_1(\cdot)$ and $\beta_2(\cdot)$ for $c = 0$ and for $c = 0.9$. We notice that the estimates behave poorly in the presence of correlation between the covariates

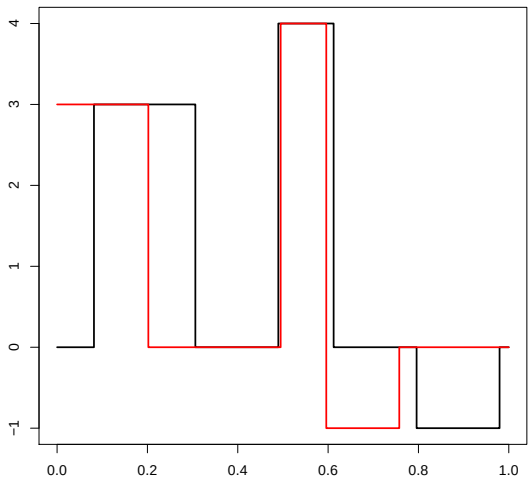


Figure 2.7: The coefficient functions $\beta_1(t)$ and $\beta_2(t)$ used to generate datasets in Section 2.3.5. *The dark (resp. red) line represents $\beta_1(t)$ (resp. $\beta_2(t)$).*

($c = 0.9$), as in a classical multiple linear regression model with scalar covariates. The investigation of the Plots (a), (b) and (c), (d) Figure 2.10 leads us to almost the same remarks as in the previous paragraph. When the cross-covariance between the covariates increases, the estimates become less accurate. We notice additionnaly that the posterior distributions are flatter for $c = 0.9$ than for $c = 0$, hence the estimates have a higher variability when there is an important cross-covariance between the covariates.

2.4 Application to the Black Périgord Truffle Dataset

We apply the Bliss method on a dataset to predict the amount of production of black truffles given the rainfall curves. The black Périgord truffle (*Tuber Melanosporum Vitt.*) is one of the most famous and valuable edible mushrooms, because of its excellent aromatic and gustatory qualities. It is the fruiting body of a hypogeous Ascomycete fungus, which grows in ectomycorrhizal symbiosis with oak species or hazelnut trees in Mediterranean conditions. Modern truffle cultivation involves the plantation of orchards with tree seedlings inoculated with *Tuber Melanosporum*. The planted orchards could then be viewed as ecosystems that should be managed in order to favour the formation and the growth of truffles. The formation begins in late winter with the germination of haploid spores released by mature ascocarps. Tree roots are then colonised by haploid mycelium to form ectomycorrhizal symbiotic associations. Induction of the fructification (sexual reproduction) occurs in May or June (the smallest truffles have been observed in mid-June). Then the young truffles grow during summer months and are mature between the middle of November and the middle of March (harvest season). The production of truffles should thus be sensitive to climatic conditions throughout the entire year (Le Tacon et al., 2014). However, to our knowledge few studies focus on the influence of rainfall or irrigation during the entire year (Demerson and Demerson, 2014; Le Tacon et al., 2014). Our aim is therefore to investigate the influence of rainfall throughout the

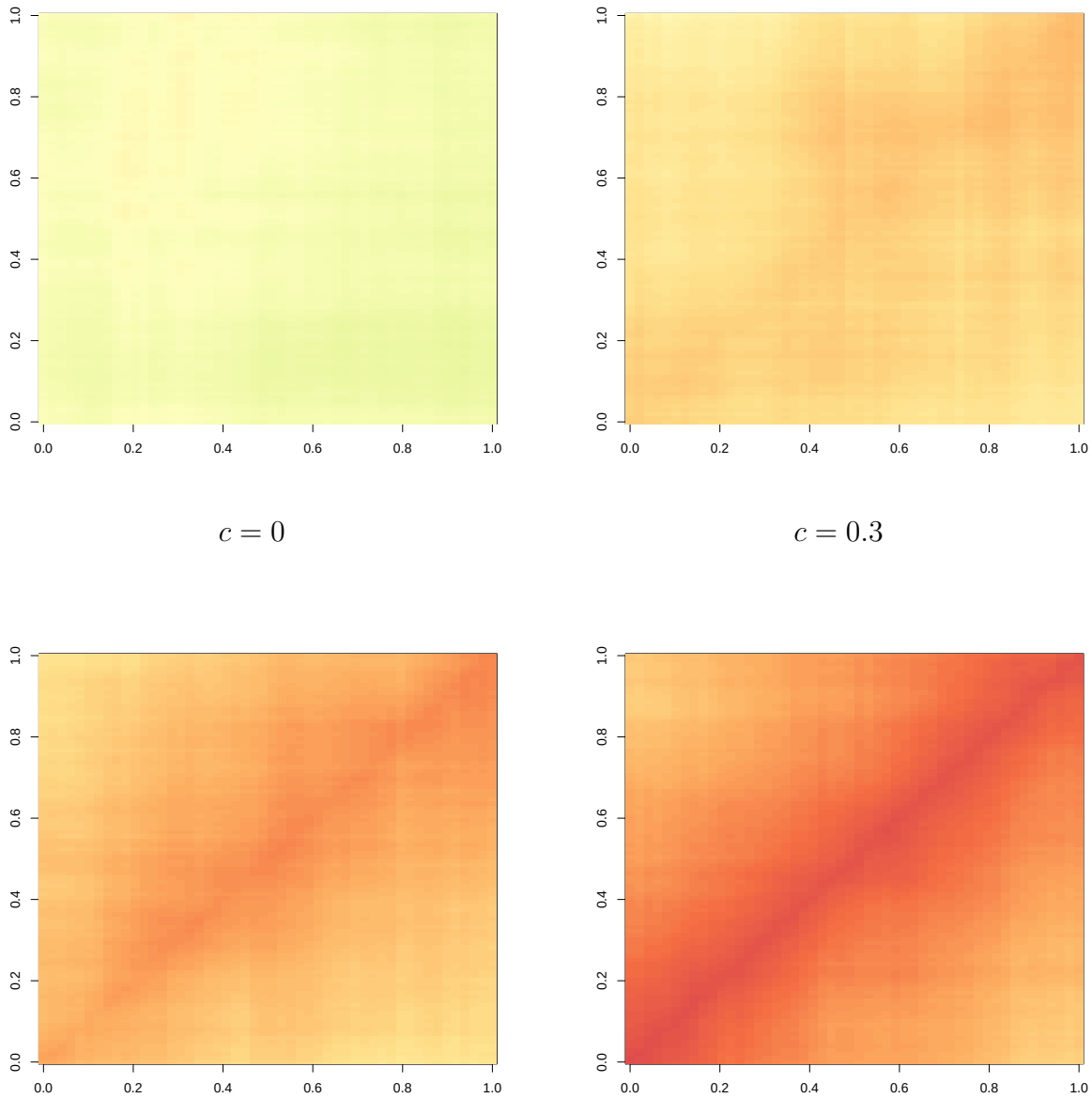


Figure 2.8: Cross-covariance matrix between the curves $x_{i1}(\cdot)$ and $x_{i2}(\cdot)$ detailed in Section 2.3.5 for different values of c. *Each point (t, t') represents the cross-covariance between $x_{i1}(t)$ and $x_{i2}(t')$. Red (resp. yellow) represents high (resp. low) cross-covariance.*

entire year on the production of black truffles. Knowing this influence could lead to better management of the orchards, to a better understanding of the sexual reproduction, and to a better understanding of the effects of climate change. Indeed, concerning sexual reproduction, Le Tacon et al. (2014, 2016) made the assumption that climatic conditions could be critical for the initiation of sexual reproduction throughout the development of the mitospores expected to occur in late winter or spring. Concerning climate change, its consequences on the geographic distribution of truffles is of interest (see Splivallo et al., 2012 or Büntgen et al., 2011, among others).

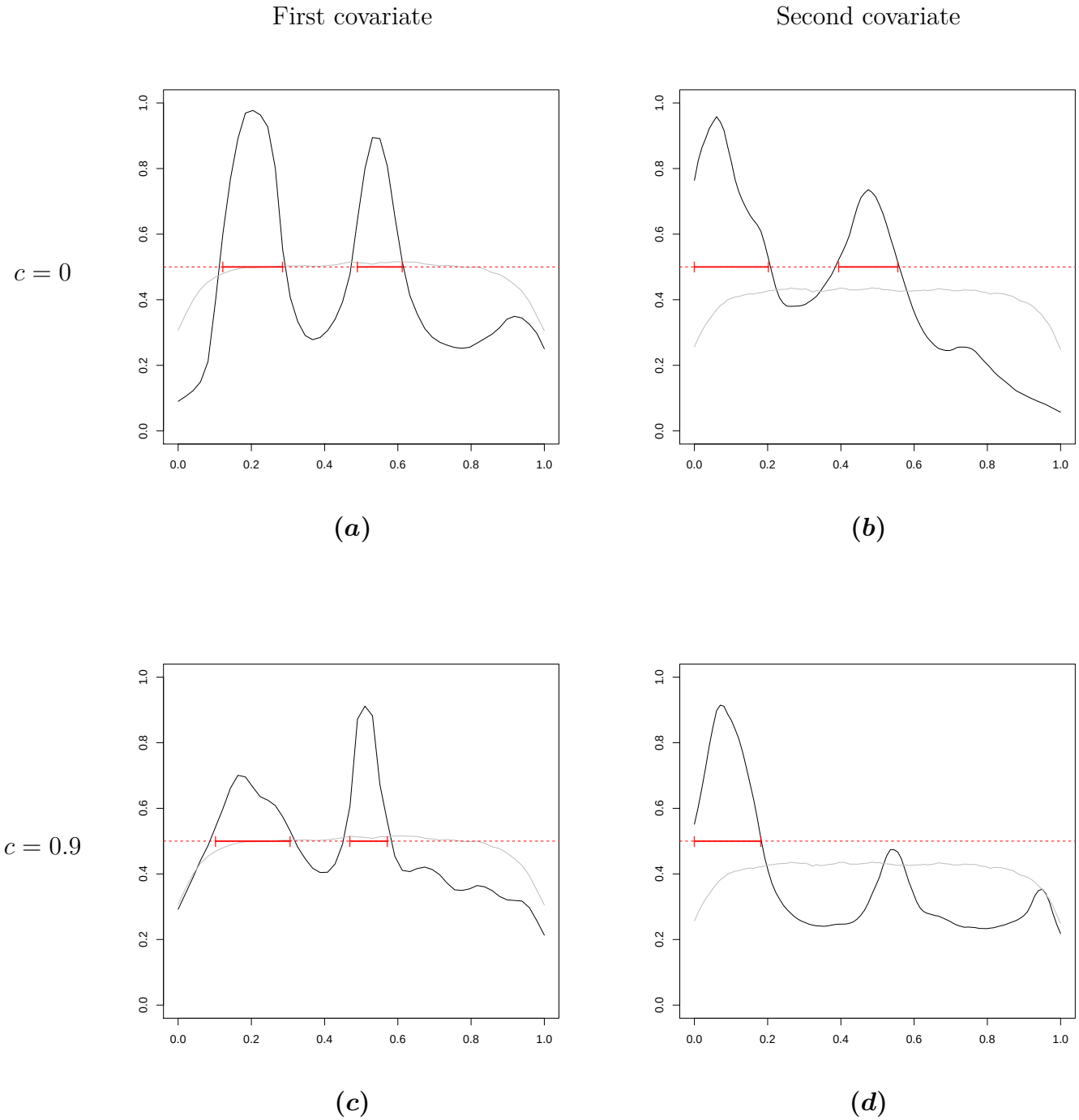


Figure 2.9: Prior (in gray) and posterior (in black) probabilities of being in the support for $c = 0$ and for $c = 0.9$. *Bayes estimates of support using Theorem 2.1 with $\gamma = 1/2$ are given in red.*

The functional covariate The analyzed data were provided by J. Demerson. They consist of the rainfall records for an orchard near Uzès (France) between 1985 and 1999, and of the production of black truffles in this orchard between 1985 and 1999. In practice, to explain the production of the year n, we take into account the rainfall between the 1st of January of the year $n-1$ and the 31st of March of the year n. Indeed, we want to take into account the whole life cycle, from the formation of new ectomycorrhizas following acospore germination during the winter preceding the harvest (year $n-1$) to the harvest of the year n. The cumulative rainfall is measured every 10 days, hence between the 1st

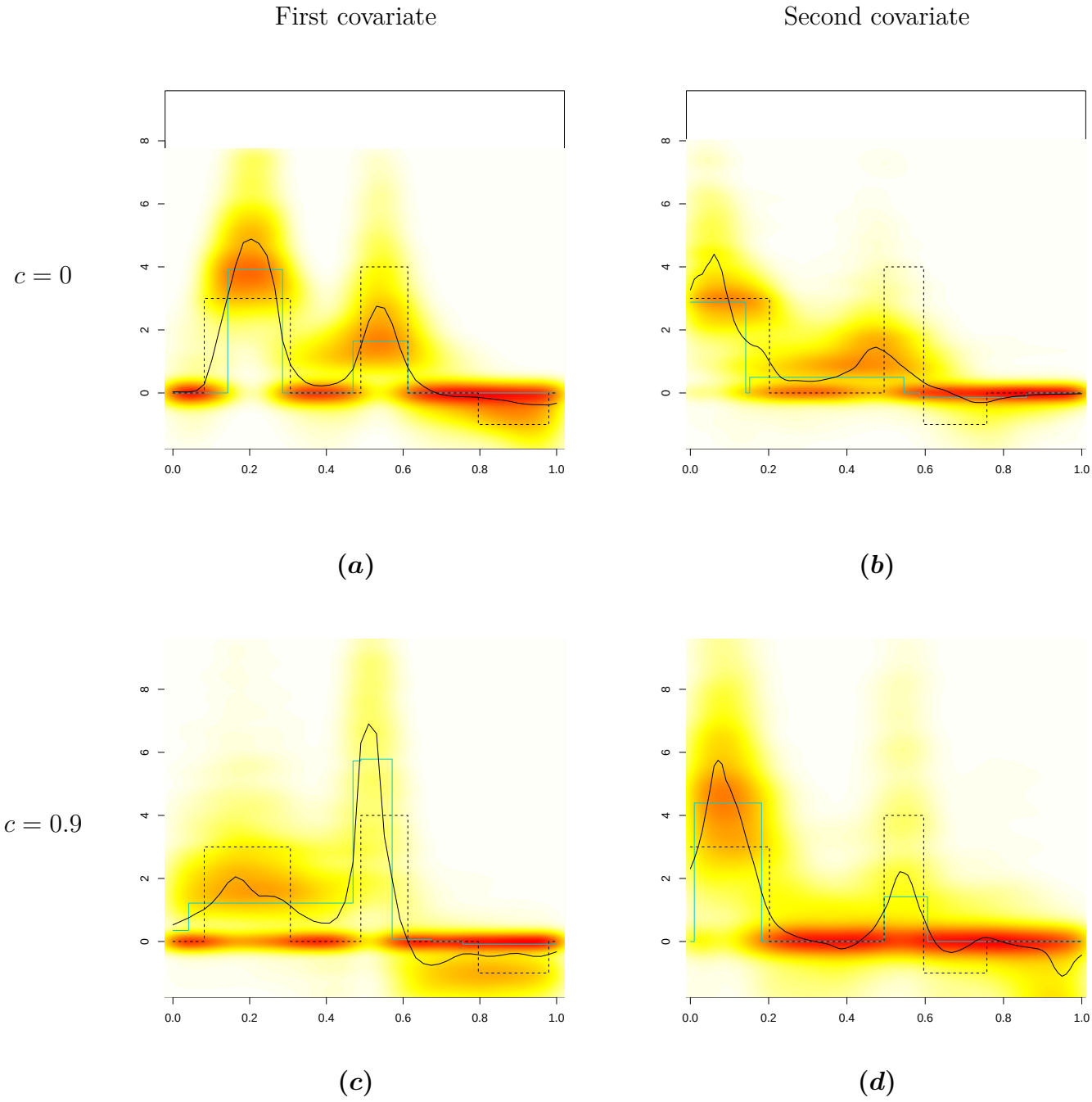


Figure 2.10: Estimates of the coefficient functions for $c = 0$ and for $c = 0.9$. *For each plot, the black dotted line is the true coefficient function, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate. The marginal posterior distributions of $\beta_\theta(t)$ are represented by using heat maps, as described in Section 2.2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities.*

of January of the year $n-1$ and the 31st of March of the year n we have the rainfall associated with 45 ten-day periods, see Figure 2.11. This dataset can be considered as reliable, as the rainfall records have been kept precisely for the orchard, and the orchard was not irrigated.

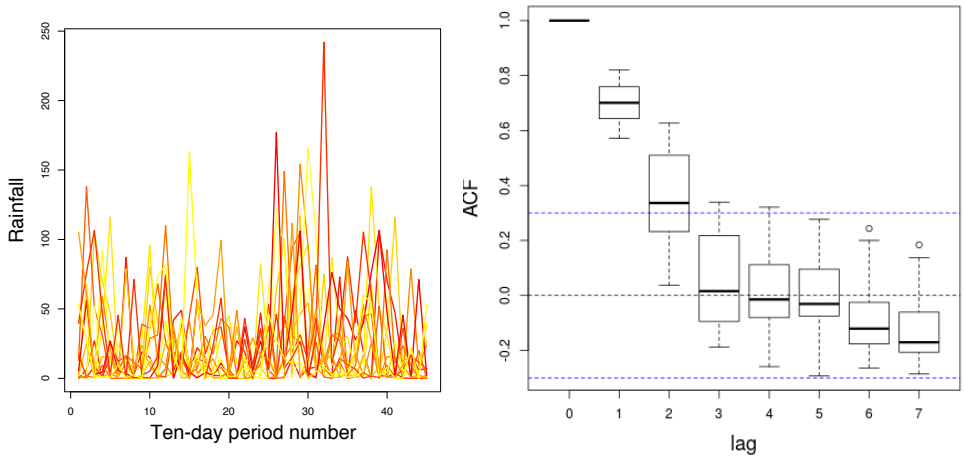


Figure 2.11: Rainfall of the Truffle datsaset. *Left:Plot shows the rainfall for each year, colour-coded by their truffle yield. Right:Autocorrelation of the 13 observed rainfall covariates, with lag in number of ten-day periods.*

Biological assumptions at stake From the literature we can spotlight the following periods of time which might influence the growth of truffles.

Period #1: Late spring and summer of year $n-1$. This is the (only) period for which all experts are unanimous in saying it has a particular effect. Büntgen et al. (2012), Demerson and Demerson (2014) or Le Tacon et al. (2014) all confirm the importance of the negative effect of summer hydric deficit on truffle production: they found it to be the most important factor influencing the production. Indeed, in summer the truffles need water to survive the high temperatures and to grow. Otherwise they can dry out and die.

Period #2: Late winter of year $n-1$, as shown by Demerson and Demerson (2014) and Le Tacon et al. (2014). Indeed, as explained in Le Tacon et al. (2014), consistent water availability in late winter could support the formation of new mycorrhizae, thus allowing a new cycle. Moreover, from results obtained by Healy et al. (2013) they made the assumption that rainfall is critical for the initiation of sexual reproduction throughout the development of mitospores, which is expected to occur in late winter or spring of the year $n-1$. This is only an assumption as the factors influencing the occurrence and the initiation of sexual reproduction are largely unknown, see Murat et al. (2013) or Le Tacon et al. (2016).

Period #3: November and December of year $n-1$, as claimed by Demerson and Demerson (2014) and Le Tacon et al. (2014). Le Tacon et al. explained that rainfall in autumn allows the growth of young truffles which have survived the summer.

Period #4: September of year $n-1$, as claimed by Demerson and Demerson (2014). Excess water in this period could be harmful to truffles. The assumption made was that in September the soil temperature is still high, so micro-organisms responsible for rot are quite active, while a wet truffle has its respiratory system disturbed and can not defend itself against these micro-organisms.

The challenge is to confirm some of these periods with Bliss, despite the small size of the dataset.

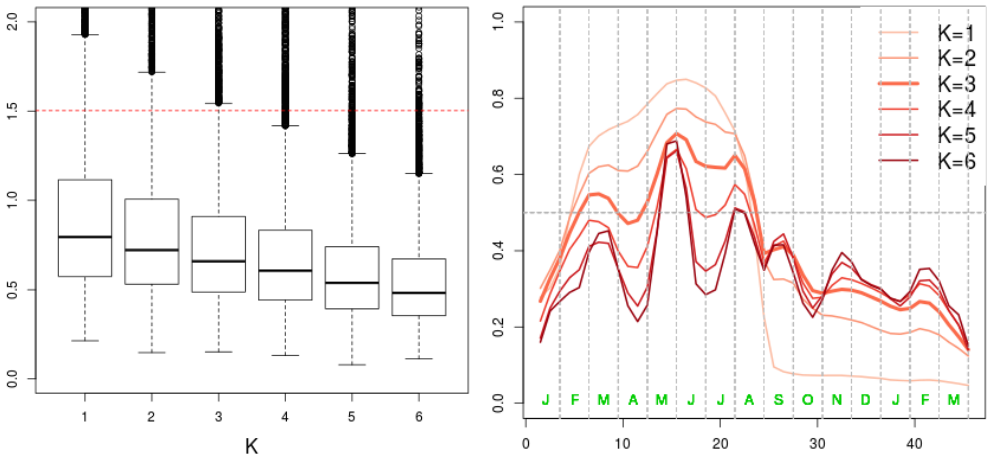


Figure 2.12: Sensitivity of Bliss to the value of K on the truffle dataset. *Left: Boxplot of the posterior distribution of the variance of the error, σ^2, compared to the variance of the output y (red dashed line). Right: Posterior probability $\alpha(t|\mathcal{D})$ for different values of K.*

Running Bliss As explained above (in Section 3.2), part of the difficulty of the inference problem comes from autocorrelation within the covariate. Figure 2.11 shows that the autocorrelation can be considered as null when the lag is 3 or more in number of ten-day periods. In other words the rainfall background presents autocorrelation within a period of time of about a month (keeping in mind that the whole history we consider lasts 15 months).

The first and maybe most important hyperparameter is K, the number of intervals in the coefficient functions from the prior. Because of the discretization of the rainfall, and the number of observations, the value of K should stay small to remain parsimonious. Because of the size of the dataset, we have set the hyperparameter a to obtain a prior probability of being in the support of about 0.5. The results are given in Figure 2.12. As can be seen on the left of this Figure, the error variance σ^2 decreases when K increases, because models of higher dimension can more easily fit the data. The main question is when do they overfit the data? In this case, the Bayesian Information Criterion selects the model with $K = 2$ intervals (see the supplementary materials). Given the small number of observations ($n = 25$), the values of BIC have to be carefully interpreted. Otherwise, looking at the right panel of Figure 2.12, we can consider how the posterior probability $\alpha(t|\mathcal{D})$ depends on the value of K and choose a reasonable value. First, for $K = 1$ or 2, the posterior probability is high during a first long period of time until August of year $n-1$ and falls to much lower values after that. Thus, these small values of K provide a rough picture of dependency. Secondly, for $K = 4, 5$ or 6, the posterior probability $\alpha(t|\mathcal{D})$ varies between 0.2 and 0.7 and shows doubtful variations after November of year $n-1$ and other strong variations during the summer of year $n-1$ that are also doubtful. Hence we decided to rely on $K = 3$ although this choice is rather subjective.

Conclusions on the truffle dataset We begin by noting that about half of the variance of the output (the amount of production of truffles) is explained by the rainfall given the posterior distribution of σ^2 in the left panel of Figure 2.12. The support estimate $\widehat{S}_{0.5}(\mathcal{D})$ with $K = 3$ is composed of two disjoint intervals: a first one from May of year $n-1$ to the second ten-day period of August with the highest posterior probability,

and a second one from the third ten-day period of February of year $n-1$ to the end of March of year $n-1$ with a smaller posterior probability. Therefore, as far as we can tell from this analysis, Periods #1 and #2 are validated by the data. Period #3 cannot be validated although the posterior probability $\alpha(t|\mathcal{D})$ presents small bumps around these periods of time for the highest values of K. For $K = 3$, the value of $\alpha(t|\mathcal{D})$ stays around 0.3 on Period #3. Finally, regarding Period #4, we can see a small bump on the curve $\alpha(t|\mathcal{D})$ around this period of time even for $K = 3$, but the highest value of the posterior probability on this period is about 0.4. Hence we choose to remain undecided on Period #4.

2.5 Conclusion

In this paper, we have provided a full Bayesian methodology to analyse linear models with time-dependent functional covariates. The main purpose of our study was to estimate the support of the coefficient function to search for the periods of time which influence the outcome the most. We rely on piecewise constant coefficient functions to set the prior, which has four benefits. The first benefit is parsimony of the Bliss model, which turns two thirds of the parameter dimension to the estimation of the support. The second benefit with our Bayesian setting that begins by defining the support is that we can rely on the Ridge-Zellner prior to handle the autocorrelation within the functional covariate. This fact sets Bliss apart from Bayesian methods relying on spike-and-slab prior to handle sparsity. The third benefit is avoiding cross-validation to tune the internal parameters of the method. Indeed, cross-validation methods optimize the performance regarding the model's predictive power, and not the accuracy of the support estimate. Last but not least, the fourth benefit is the ability to compute the posterior probability that a given date is in the support, $\alpha(t|\mathcal{D})$, whose value gives a clear hint on the reliability of the support estimate. Nevertheless a serious limitation of our Bayesian model is that it becomes difficult to handle a d-dimensional functional covariates, for $d > 1$. Indeed the shape of the support of a function of more than one variable is much more complex than a union of intervals and cannot be easily modelled in a nonparametric, but parsimonious manner.

We have provided numerical results regarding the power of Bliss on a bunch of synthetic datasets as well as a dataset studying the black Périgord truffle. We have shown by presenting some of these examples in detail how we can interpret the results of Bliss, in particular how we can rely on the posterior probabilities $\alpha(t|\mathcal{D})$ or the heatmap of posterior distribution of the coefficient function to assess the reliability of our estimates. Bliss provides two main outputs: first an estimate of the support of the coefficient function without targeting the coefficient function, and second a trade-off between support estimate and coefficient function estimate through the stepwise estimate of Proposition 2.3. Moreover our prior can straightforwardly be encompassed into a linear model with other functional or scalar covariates.

2.6 Appendices

2.6.1 Theoretical Results

Proof of Theorem 2.1

Without loss of generality we can assume that $\mathcal{T} = [0, 1]$. We begin the proof with the following lemma whose simple proof is left to the reader.

Lemma 2.4. *Set* $\psi^*(\gamma, \alpha) = \min\{\gamma(1-\alpha)\,;\,(1-\gamma)\alpha\}$ *for any* $\alpha, \gamma \in [0, 1]$. *We have*

$$\psi^*(\gamma, \alpha) = \begin{cases} \gamma(1-\alpha) & \text{if } \gamma \leq \alpha, \\ (1-\gamma)\alpha & \text{if } \gamma \geq \alpha. \end{cases}$$

Remember that the posterior loss we optimise is given in (2.12), where S is any Borel subset of $\mathcal{T} = [0, 1]$. Using Fubini's theorem (for non-negative functions) and the definition of $\alpha(t|\mathcal{D})$ given in (2.10), we have

$$\begin{aligned} \int_{\Theta_K} L_\gamma(S, S_\theta)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta &= \gamma \int_0^1 \int_{\Theta_K} \mathbf{1}\{t \in S \setminus S_\theta\}\pi_K(\theta|\mathcal{D})\mathrm{d}\theta\mathrm{d}t \\ &\quad + (1-\gamma)\int_0^1 \int_{\Theta_K} \mathbf{1}\{t \in S_\theta \setminus S\}\pi_K(\theta|\mathcal{D})\mathrm{d}\theta\mathrm{d}t \\ &= \int_0^1 \psi_S\Big(t, \gamma, \alpha(t|\mathcal{D})\Big)\mathrm{d}t \end{aligned} \tag{2.22}$$

where, for all $\alpha \in [0, 1]$ we have set

$$\psi_S(t, \gamma, \alpha) = \mathbf{1}\{t \in S\}\gamma\Big(1-\alpha\Big) + \mathbf{1}\{t \notin S\}(1-\gamma)\alpha.$$

Now, whatever the set S, $\psi_S(t, \gamma, \alpha) \geq \psi^*(\gamma, \alpha)$. Reporting this bound in (2.22) yields

$$\int_{\Theta_K} L_\gamma(S, S_\theta)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta \geq \int_0^1 \psi^*\Big(\gamma, \alpha(t|\mathcal{D})\Big)\mathrm{d}t$$

whatever the Borel set S. Moreover, this inequality is an equality if and only if the Borel set S is chosen so that, for almost all $t \in [0, 1]$, $\psi_S\Big(t, \gamma, \alpha(t|\mathcal{D})\Big) = \psi^*\Big(\gamma, \alpha(t|\mathcal{D})\Big)$. Using Lemma 2.4, the last condition is equivalent to saying that for almost all $t \in [0, 1]$, either $\alpha(t|\mathcal{D}) = \gamma$ or $\Big(t \in S \iff \gamma \leq \alpha(t|\mathcal{D})\Big)$. This concludes the proof of Theorem 2.1. □

Proof of Proposition 2.2

Obviously, $\widehat{\beta}_{L^2}(\cdot)$ minimizes

$$\int \int_{\mathcal{T}} \left(\beta_\theta(t) - d(t)\right)^2 \mathrm{d}t \; \pi_K(\theta|\mathcal{D})\mathrm{d}\theta = \int_{\mathcal{T}} \int \left(\beta_\theta(t) - d(t)\right)^2 \pi_K(\theta|\mathcal{D})\mathrm{d}\theta \; \mathrm{d}t$$

because it need to optimize $\int \left(\beta_\theta(t) - d(t)\right)^2 \pi_K(\theta|\mathcal{D})\mathrm{d}\theta$ for all $t \in \mathcal{T}$. It remains to show that $\widehat{\beta}_{L^2}(\cdot) \in L^2(\mathcal{T})$. We have

$$\begin{aligned}
\|\widehat{\beta}_{L^2}(\cdot)\|^2 &= \int_{\mathcal{T}} \left(\int \beta_\theta(t)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta \right)^2 \mathrm{d}t \\
&= \int_{\mathcal{T}} \iint \beta_\theta(t)\beta_{\theta'}(t)\ \pi_K(\theta|\mathcal{D})\pi_K(\theta'|\mathcal{D})\mathrm{d}\theta\mathrm{d}\theta'\mathrm{d}t \\
&= \iint \int_{\mathcal{T}} \beta_\theta(t)\beta_{\theta'}(t)\mathrm{d}t\ \pi_K(\theta|\mathcal{D})\pi_K(\theta'|\mathcal{D})\mathrm{d}\theta\mathrm{d}\theta' \\
&\leq \iint \|\beta_\theta(\cdot)\|\|\beta_{\theta'}(\cdot)\|\ \pi_K(\theta|\mathcal{D})\pi_K(\theta'|\mathcal{D})\mathrm{d}\theta\mathrm{d}\theta' \quad \text{with Cauchy-Schwarz inequality} \\
&\leq \left(\int \|\beta_\theta(\cdot)\|\pi_K(\theta|\mathcal{D})\mathrm{d}\theta \right)^2
\end{aligned}$$

And the last integral is finite because of the assumption. Hence $\widehat{\beta}_{L^2}(\cdot)$ is in $L^2(\mathcal{T})$. □

Proof of Proposition 2.3

First, the norm $\|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|$ is non negative, hence the set

$$\left\{ \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|,\ d(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0} \right\}$$

admits an infimum. Let m denote this infimum. We have to prove that m is actually a minimum of the above set, namely that there exists a function $d(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0}$ so that $m = \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|$.

To this end, we introduce a minimizing sequence $\{d_n(\cdot)\}$ and we will show that one of its subsequences admits a limit within $\mathcal{E}^{\varepsilon}_{K_0}$. Let $d_n(\cdot)$ be so that

$$m = \inf \left\{ \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|,\ d(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0} \right\} \leq \|d_n(\cdot) - \widehat{\beta}_{L^2}(\cdot)\| \leq m + 2^{-n}. \tag{2.23}$$

The step function $d_n(\cdot)$ can be written as

$$d_n(t) = \sum_{k=1}^{L} \alpha_{k,n} \mathbf{1}\{t \in (a_{k,n},\ b_{k,n})\}$$

where the set $(a_{k,n},\ b_{k,n})$, $k = 1, \ldots, L$ forms a partition of $[0, 1]$. Note that their number L does not depend on n because all $d_n(\cdot)$ lie in $\mathcal{E}_{K_0}$ for some fixed value of K_0, and we can always choose $L = 2K_0 + 1$. Moreover, because $d_n(\cdot)$ is in $\mathcal{F}^{\varepsilon}$, we can assume that

$$b_{k,n} - a_{k,n} \geq \varepsilon, \quad \text{for all } k, n. \tag{2.24}$$

Now the sequence $\{a_{1,n}\}_n$ has its elements in the compact interval $\mathcal{T}$ hence we extract a subsequence (still denoted $\{a_{1,n}\}_n$) which converges to an element $a_{1,\infty}$ of $\mathcal{T}$. Likewise,

by the Bolzano-Weierstrass theorem we extract subsequences $2L$ times and we can assume that all sequences $\{a_{1,n}\}_n, \ldots, \{a_{L,n}\}_n$, $\{b_{1,n}\}_n, \ldots, \{b_{L,n}\}_n$ are convergent, and that

$$a_{k,\infty} = \lim_{n\to\infty} a_{k,n}, \quad b_{k,\infty} = \lim_{n\to\infty} b_{k,n}, \quad \text{and} \quad b_{k,\infty} - a_{k,\infty} \geq \varepsilon, \quad k = 1, \ldots, L$$

where the last inequalities come from (2.24).

The function $d_n(\cdot)$ is bounded (in L^2-norm):

$$\|d_n(\cdot)\| \leq \|\widehat{\beta}_{L^2}(\cdot)\| + \|d_n(\cdot) - \widehat{\beta}_{L^2}(\cdot)\| \leq R + m + 1$$

with (2.23), where $R = \|\widehat{\beta}_{L^2}(\cdot)\|$ with $R < \infty$ because $\widehat{\beta}_{L^2}(\cdot) \in L^2(\mathcal{T})$ by Proposition 2.14. Moreover

$$\|d_n(\cdot)\|^2 = \sum_{k=1}^{L} \alpha_{k,n}^2 \Big(b_{k,n} - a_{k,n}\Big) \geq \varepsilon \sum_{k=1}^{L} \alpha_{k,n}^2.$$

Hence, each sequence $\{\alpha_{1,n}\}_n, \ldots, \{\alpha_{L,n}\}_n$ is bounded. Thus, by further extracting sub-subsequences, we can assume that, for $k = 1, \ldots, L$,

$$\lim_{n\to\infty} \alpha_{k,n} = \alpha_{k,\infty}$$

Finally, by setting

$$d_\infty(\cdot) = \sum_{k=1}^{L} \alpha_{k,\infty} \mathbf{1}\{t \in (a_{k,\infty},\ b_{k,\infty})\}$$

we can easily prove that $d_n(\cdot)$ tends to $d_\infty(\cdot)$ in L^2-norm and that $d_\infty(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0}$. And, with (2.23)

$$m = \|d_\infty(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|$$

which concludes the proof of (i).

The proof of (ii) is quite obvious because if $d \in \mathcal{E}^{\varepsilon}_{K_0}$,

$$\begin{aligned}\int L^{\varepsilon}_{K_0}\Big(d(\cdot), \beta_\theta(\cdot)\Big)\, \pi_K(\theta|\mathcal{D})\, \mathrm{d}\theta &= \|d(\cdot)\|^2 - 2 < d(\cdot), \int \beta_\theta(\cdot)\, \pi_K(\theta|\mathcal{D})\, \mathrm{d}\theta > + \int \|\beta_\theta(\cdot)\|^2\, \pi_K(\theta|\mathcal{D})\, \mathrm{d}\theta \\ &= \left\| d(\cdot) - \int \beta_\theta(\cdot) \pi_K(\theta|\mathcal{D})\, \mathrm{d}\theta \right\|^2 \\ &\quad + \int_{\mathcal{T}} \left[\int \beta_\theta(t)^2\, \pi_K(\theta|\mathcal{D})\, \mathrm{d}\theta - \left(\int \Big(\beta_\theta(t)\, \pi_K(\theta|\mathcal{D})\, \mathrm{d}\theta \Big)^2 \right] \end{aligned}$$

□

Topological Properties of $\mathcal{E}_K$

Proposition 2.5. *Let $K \geq 1$.*

(i) The convex hull of $\mathcal{E}_K$ is $\mathcal{E}$.

(ii) Under the $L^2(\mathcal{T})$-topology, the closure of $\mathcal{E}$ is $L^2(\mathcal{T})$.

Proof. The result of *(ii)* is rather classical, see, e.g., Rudin (1986). The convex hull of $\mathcal{E}_K$ includes any step function. Indeed, any convex combination of step functions of $\mathcal{E}_K$ belongs to $\mathcal{E}$. Moreover, $\mathcal{E}$ is convex because it is a linear space. Hence claim *(i)* is proven. □

For a given K, the set of functions $\mathcal{E}_K$ is not suitable to define a projection of $\hat{\beta}_{L^2}(\cdot)$. Indeed, let $\{d_n(\cdot)\}$ be a minimizing sequence of the set $\left\{\|d(\cdot) - \hat{\beta}_{L^2}(\cdot)\|,\ d(\cdot) \in \mathcal{E}_K(\cdot)\right\}$, such that

$$m = \inf\left\{\|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|,\ d(\cdot) \in \mathcal{E}_K\right\} \leq \|d_n(\cdot) - \widehat{\beta}_{L^2}(\cdot)\| \leq m + 2^{-n}.$$

Knowing that $\hat{\beta}_{L^2}(\cdot)$ and $d_n(\cdot)$ belong to $L^2(\mathcal{T})$ for all n, we have

$$d_n(.) \in \mathcal{E}_K \cap \mathcal{B}_{L^2}(R + m + 1), \qquad \text{for all } n,$$

where $\mathcal{B}_{L^2}(r)$ is the $L^2(\mathcal{T})$-ball of radius r around the origin. Note that $\mathcal{E}_K \cap \mathcal{B}_{L^2}(R+m+1)$ is not a compact set, for instance it is not possible to extract a convergent subsequence from $d_n(t) = \sqrt{n}\,\mathbf{1}\{t \in [0, \frac{1}{n}]\}$. Hence it is not possible to extract a subsequence of $\{d_n(\cdot)\}$ which converges to a $d_\infty(\cdot) \in \mathcal{E}_K$ so that $\|d(\cdot) - \hat{\beta}_{L^2}(\cdot)\| = m$.

2.6.2 Details of the Implementations

Gibbs algorithm and Full conditional distributions for a single functional covariate

The full conditional distributions for the Gibbs Sampler in Section 2.2.7 are as follows,

$$\begin{aligned}
\mu, b | y, \sigma^2, m, \ell &\sim \mathcal{N}_{K+1}\left((\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\underline{\mathrm{x}}y\ ,\ \sigma^2(\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\right), \\
\sigma^2 | y, \mu, b, m, \ell &\sim \Gamma^{-1}\left(\frac{n+K+1}{2}, \frac{1}{2}\mathrm{RSS} + \frac{1}{2}(\mu, b)^T\underline{\mathrm{V}}^{-1}(\mu, b)\right), \\
\pi\left(m_k | y, \mu, b, \sigma^2, m_{-k}, \ell\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(b | m, \ell, \sigma^2) \\
\pi\left(\ell_k | y, \mu, b, \sigma^2, m, \ell_{-k}\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(\ell_k) \times \pi(b | m, \ell, \sigma^2)
\end{aligned}$$

where $\mathrm{RSS} = \|y - \mu\mathbf{1}_n - x_{.}(\mathcal{I}_{.})b\|^2$, $\underline{\mathrm{x}} = \left(\mathbf{1}_n \ \middle|\ x_{.}(\mathcal{I}_{.})\right)$, and

$$\underline{\mathrm{V}} = \begin{pmatrix} v_0^{-1} & 0 \\ 0 & n^{-1}\Big(G + v\lambda_{\max}(G) I_K\Big) \end{pmatrix},$$

where $G = x_{.}(\mathcal{I}_{.})^T x_{.}(\mathcal{I}_{.})$. The full conditional distributions for the hyperparameters m_k and ℓ_k are unusual distributions. As the covariate curves $x_i(\cdot)$ are observed on a grid $\mathcal{T}_G = (t_j)_{j=1,\ldots,p}$, we consider that m_k belongs to $\mathcal{T}_G$ and ℓ_k is defined so that $m_k \pm \ell_k \in \mathcal{T}_G$. Thus, the number of possible values for m_k and ℓ_k is finite and the full conditional distributions of m_k and ℓ_k are easily computable.

Gibbs Algorithm and Full Conditional Distributions for q Functional Covariates

Remenber $K = \sum_{j=1}^{q} K_j$. We denote

- $b_j = (b_{1j}, \ldots, b_{K_j j})$ and $b = (b_1, \ldots, b_q)$,
- $m_j = (m_{1j}, \ldots, m_{K_j j})$ and $m = (m_1, \ldots, m_q)$,
- $\ell_j = (\ell_{1j}, \ldots, \ell_{K_j j})$ and $\ell = (\ell_1, \ldots, \ell_q)$.

The full conditional distributions are

$$\begin{aligned}
\mu, b | y, \sigma^2, m, \ell &\sim \mathcal{N}_{K+1}\left((\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\underline{\mathrm{x}}y \ ,\ \sigma^2(\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\right), \\
\sigma^2 | y, \mu, b, m, \ell &\sim \Gamma^{-1}\left(\frac{n+K+1}{2}, \frac{1}{2}\mathrm{RSS} + \frac{1}{2}(\mu, b)^T\underline{\mathrm{V}}^{-1}(\mu, b)\right), \\
\pi\left(m_{kj} | y, \mu, b, \sigma^2, m_{-(kj)}, \ell\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(b | m, \ell, \sigma^2) \\
\pi\left(\ell_{kj} | y, \mu, b, \sigma^2, m, \ell_{-(kj)}\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(\ell_{kj}) \times \pi(b | m, \ell, \sigma^2)
\end{aligned}$$

where $\mathrm{RSS} = \left\| y - \mu\mathbf{1}_n - \sum_{j=1}^{q} x_{\cdot j}(\mathcal{I}_{\cdot j})b_j \right\|^2$, $\underline{\mathrm{x}} = \left(\mathbf{1}_n \ \middle|\ x_{\cdot 1}(\mathcal{I}_{\cdot 1}) \ \middle|\ \ldots \ \middle|\ x_{\cdot q}(\mathcal{I}_{\cdot q})\right)$ and

$$\underline{\mathrm{V}} = \begin{pmatrix}
v_0^{-1} & 0 & \ldots & 0 \\
0 & n^{-1}\left(G_1 + v\lambda_{\max}(G_1)I_{K_1}\right) & & 0 \\
\vdots & & \ddots & \\
0 & 0 & & n^{-1}\left(G_{K_q} + v\lambda_{\max}(G_{K_q})I_{K_q}\right)
\end{pmatrix},$$

where $G_j = x_{\cdot j}(\mathcal{I}_{\cdot j})^T x_{\cdot j}(\mathcal{I}_{\cdot j})$.

Simulated Annealing Algorithm

We give in this section the details of the Simulated Annealing algorithm we use. In presence of more than one functional covariate, the following algorithm is used to determine the estimate of each coefficient function one by one.

Let $\tilde{\Theta}_{K_0} = \bigotimes_{K=1}^{K_0}\left(K, \Theta_K\right)$ where Θ_K is the space of all $\theta = (b_1, \ldots, b_K, m_1, \ldots, m_K, \ell_1, \ldots, \ell_K)$ and let the function $C(d(\cdot)) = \left\| d(\cdot) - \hat{\beta}_{L^2}(\cdot) \right\|^2$.

Algorithm : *Simulated Annealing*

- Initialize: a deterministic decreasing schedule of temperature $(\tau_i)_{i=1,\ldots,N_{\mathrm{SANN}}}$, a value of K_0 and an initial vector $(K_{(0)}, \theta_{(0)}) \in \tilde{\Theta}_{K_0}$.
- Compute the function $\beta_{(0)}(t)$ from $(K_{(0)}, \theta_{(0)})$.

- Repeat for i from 1 to N_{SANN} :
  - Choose randomly a move from $(K_{(i-1)}, \theta_{(i-1)})$ to (K', θ') among :
    1. propose a new ${b_k}'$ for an arbitrary $k \leq K_{(i-1)}$,
    2. propose a new m'_k for an arbitrary $k \leq K_{(i-1)}$,
    3. propose a new ℓ'_k for an arbitrary $k \leq K_{(i-1)}$,
    4. propose to append a new interval (b', m', ℓ') or
    5. propose to drop out an interval (b_k, m_k, ℓ_k) for an arbitrary $k \leq K_{(i-1)}$.
  - Compute the function $\beta'(t)$ from the proposal (K', θ').
  - Compute the acceptance ratio

$$\alpha = \min\left\{1, \exp\left(\frac{C(\beta'(\cdot)) - C\Big(\beta_{(i-1)}(\cdot)\Big)}{\tau_i}\right)\right\}.$$

  - Draw u from $\text{Unif}(0, 1)$.
  - If $u < \alpha$, $(K_{(i)}, \theta_{(i)}) = (K', \theta')$ (move accepted), else $(K_{(i)}, \theta_{(i)}) = (K_{(i-1)}, \theta_{(i-1)})$ (move rejected).
  - Compute the function $\beta_{(i)}(t)$ from $(K_{(i)}, \theta_{(i)})$.
- Return the iteration $(K_{(i)}, \theta_{(i)})$ minimizing the criteria $C(.)$.

For the schedule of temperature, we use by default a logarithmic schedule (see Bélisle, 1992), which is given for each iteration i by

$$\text{Te}/\log\left((i-1)+e\right), \tag{2.25}$$

where Te is a parameter to calibrate and corresponds to the initial temperature. The result of the Simulated Annealing algorithm is sensitive to the scale of Te and it is quite difficult to find a suitable a priori value. For example, if the initial temperature is too small, almost all the proposed moves are rejected during the algorithm. On the other hand, if it is too large, they are almost all accepted. So, we run the algorithm a few times and each time Te is determined with respect to the previous runs. For instance, if for a run the moves are always rejected or always accepted, the initial temperature for the next run is accordingly adjusted. Only 2 or 3 runs are necessary to find a suitable scale of Te in our applications.

2.6.3 Computational Time

In this Section, we provide the computational time of the 3 algorithms used in this paper:

1. a Gibbs sampler, for which the full conditional distributions are given in Section 2.6.2,

2. an algorithm, denoted *density estimation*, which computes the heat map described in Section 2.2.7 and

3. a simulated annealing algorithm described in Appendix 2.6.2.

The computational time of the simulated annealing algorithm is negligible ($\sim 1s$). Moreover, the computational time of the *density estimation* algorithms is around one minute and it is mainly due to use of the *kde2d* function. Therefore, below we discuss only the Gibbs sampler.

First, remember that we observe n curves for q covariates, evaluated on the regular grid $\boldsymbol{t^j} = (t_1^j, \ldots, t_p^j)$ for the j^{th} covariates. Remember also that K_j are the fixed number of intervals in (2.5) for the j^{th} dimension and $K = \sum_j^q K_j$. We give a sketch of the main steps of the Gibbs sampler algorithm we use. The Gibbs sampler consists of two major steps. Firstly, we compute the integrals $\int_{\mathcal{I}} x_{ij}(t)\mathrm{d}t$ on all possible intervals $\mathcal{I}$ for $j = 1, \ldots, q$. The intervals $\mathcal{I}$ depend on a center m and a half-length ℓ. In practice, we consider that m belongs to the grid $\boldsymbol{t}$ and we consider p possible values for ℓ. Hence, we have to compute $n \times p^2 \times q$ integrals. Secondly, we perform a loop of N iterations for which each parameter is updated with regard to full conditional distributions. At each iteration, the main required computations are:

- q Singular Value Decompositions of matrix $n \times K_j$ to compute the matrix G_j in (2.18),
- inversions of a matrix $(K+1) \times (K+1)$ and
- determinants of a matrix $(K+1) \times (K+1)$.

For one single functional covariate For the case of one single functional covariate, we use a data set described in Section 2.3.1 with different values for n and p (50, 100 and 200). We apply the Bliss method with $K = 3$ or $K = 6$ and different number iterations (10 000, 20 000 and 50 000). The computational times are given in Table 2.4.

For two functional covariates For the case of two functional covariates, we use a data set described in Section 2.3.1 with different values for n, p_1 and p_2. We vary the value of K and the number iterations as in the previous paragraph. The computational times are given in Table 2.5.

2.6.4 Model Choice by using BIC

Below, we illustrate the model choice procedure in order to fix the hyperparameter K. We apply the procedure on different simulated datasets and it is applied on the truffle dataset.

Table 2.4: *The computational time of the Gibbs sampler for different values of n, p, K and the number iterations in the one single functional covariate case.*

n	p	K	*iter*	computational time	n	p	K	*iter*	computational time
50	50	3	10 000	1.3 min	50	50	6	20 000	11.5 min
100	50	3	10 000	3.3 min	100	50	6	20 000	26.8 min
200	50	3	10 000	11.5 min	200	50	6	20 000	1.5 h
50	100	3	10 000	2.5 min	50	100	6	20 000	23.6 min
100	100	3	10 000	6.7 min	100	100	6	20 000	54.7 min
200	100	3	10 000	23.3 min	200	100	6	20 000	3 h
50	200	3	10 000	5.2 min	50	200	6	20 000	48 min
100	200	3	10 000	13.7 min	100	200	6	20 000	1.9 h
200	200	3	10 000	47.5 min	200	200	6	20 000	6 h
50	50	6	10 000	5.7 min	50	50	3	50 000	6.4 min
100	50	6	10 000	13.4 min	50	100	3	50 000	17 min
200	50	6	10 000	44.2 min	50	200	3	50 000	59.3 min
50	100	6	10 000	11.8 min	50	100	3	50 000	12.6 min
100	100	6	10 000	27.4 min	100	100	3	50 000	33.3 min
200	100	6	10 000	1.5 h	200	100	3	50 000	1.9 h
50	200	6	10 000	24.5 min	50	200	3	50 000	25.5 min
100	200	6	10 000	55.8 min	100	200	3	50 000	1.1 h
200	200	6	10 000	3 h	200	200	3	50 000	3.9 h
50	50	3	20 000	2.5 min	50	50	6	50 000	28.8 min
100	50	3	20 000	6.6 min	100	50	6	50 000	1.1 h
200	50	3	20 000	23.1 min	200	50	6	50 000	3.6 h
50	100	3	20 000	5 min	50	100	6	50 000	59.1 min
100	100	3	20 000	13.3 min	100	100	6	50 000	2.3 h
200	100	3	20 000	46.2 min	200	100	6	50 000	7.4 h
50	200	3	20 000	10.2 min	50	200	6	50 000	2 h
100	200	3	20 000	27.2 min	100	200	6	50 000	4.6 h
200	200	3	20 000	1.6 h	200	200	6	50 000	14.8 h

Table 2.5: *The computational time of the Gibbs sampler for different values of n, p_1, p_2, K and the number iterations in the two functional covariates case.*

n	p_1 and p_2	K	*iter*	computational time	n	p_1 and p_2	K	*iter*	computational time
50	50	3	10 000	6 min	50	50	6	20 000	59 min
100	50	3	10 000	13.7 min	100	50	6	20 000	2 h
200	50	3	10 000	44 min	200	50	6	20 000	5.8 h
50	100	3	10 000	12.2 min	50	100	6	20 000	2.1 h
100	100	3	10 000	27.7 min	100	100	6	20 000	4.1 h
200	100	3	10 000	1.5 h	200	100	6	20 000	11.9 h
50	200	3	10 000	25.1 min	50	200	6	20 000	4.3 h
100	200	3	10 000	56.4 min	100	200	6	20 000	8.3 h
200	200	3	10 000	3 h	200	200	6	20 000	24 h
50	50	6	10 000	29.8 min	50	50	3	50 000	30.6 min
100	50	6	10 000	59 min	50	100	3	50 000	1.2 h
200	50	6	10 000	2.9 h	50	200	3	50 000	3.7 h
50	100	6	10 000	1 h	50	100	3	50 000	1 h
100	100	6	10 000	2 h	100	100	3	50 000	2.3 h
200	100	6	10 000	5.9 h	200	100	3	50 000	7.4 h
50	200	6	10 000	2.2 h	50	200	3	50 000	2.1 h
100	200	6	10 000	4.2 h	100	200	3	50 000	4.6 h
200	200	6	10 000	12 h	200	200	3	50 000	14.8 h
50	50	3	20 000	11.8 min	50	50	6	50 000	2.5 h
100	50	3	20 000	27.4 min	100	50	6	50 000	4.9 h
200	50	3	20 000	1.5 h	200	50	6	50 000	14.6 h
50	100	3	20 000	24.3 min	50	100	6	50 000	5.1 h
100	100	3	20 000	55.6 min	100	100	6	50 000	10.1 h
200	100	3	20 000	3 h	200	100	6	50 000	29.6 h
50	200	3	20 000	49.9 min	50	200	6	50 000	10.7 h
100	200	3	20 000	1.9 h	100	200	6	50 000	20.7 h
200	200	3	20 000	6 h	200	200	6	50 000	60.2 h

Simulation study

We simulate four kinds of datasets to evaluate the performance of BIC in different situations.

First, we show the results of BIC when the interval number of the true coefficient function varies. Therefore, we simulate datasets with the coefficient functions given in Figure 2.13 for which the true K varies from 1 to 5. The values of BIC are given in Figure 2.14. When the true K is 1 or 2, BIC selects the true model. Otherwise, when the true K is greater than 3, it underestimates K which can be due to the small sample size ($n = 100$).

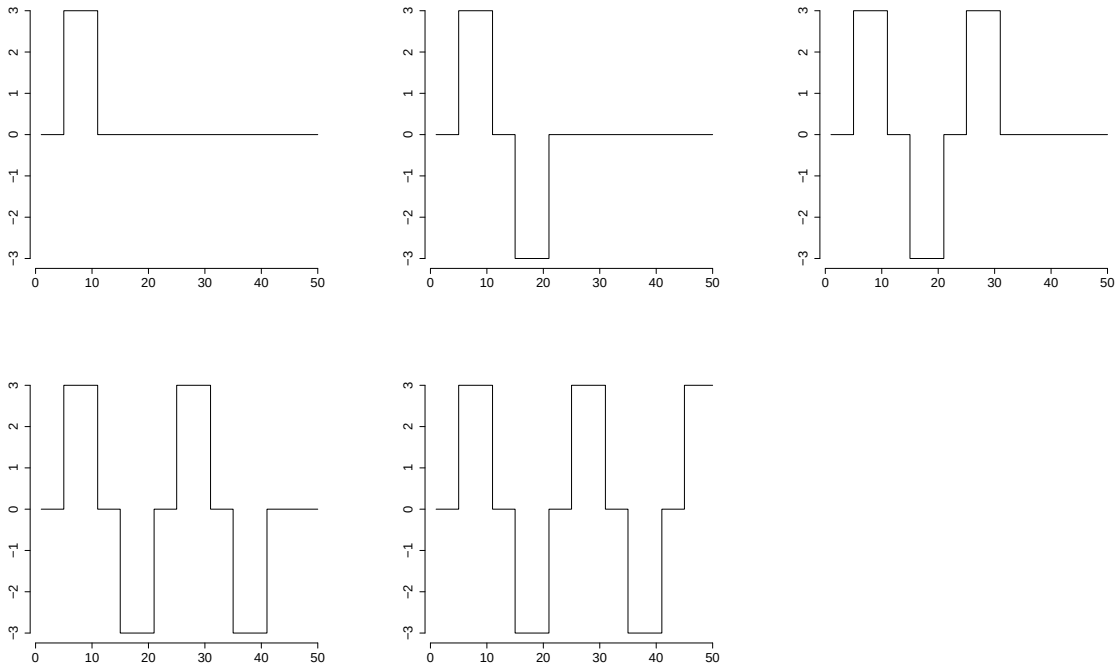


Figure 2.13: Coefficient functions with different numbers of intervals, used to simulate datasets.

Second, we illustrate the variation of BIC when the autocorrelation of the curves $x_i(\cdot)$ increases, i.e. ζ decreases (see Section 3.1). We simulate datasets with $\zeta = 1, 0.8, 0.6, 0.4$ and 0.2, with $n = 100$ and the true coefficient function is the third plot given in Figure 2.13. Plot (a) Figure 2.15 shows the values of BIC for different values of ζ. When the autocorrelation is high (blue and light blue lines), BIC underestimates K, since in this example the true K is 3. The results are the same if the true coefficient function is the Smooth function described in Section 3.1 of the main paper, see Plot (b) Figure 2.15.

Third, we illustrate the variation of BIC when the sample size n increases. Figure 2.16 shows the values of BIC for $n = 50, 100, 200$ and 300. As expected, when the sample size increases, BIC is more accurate for selecting the true model $K = 3$.

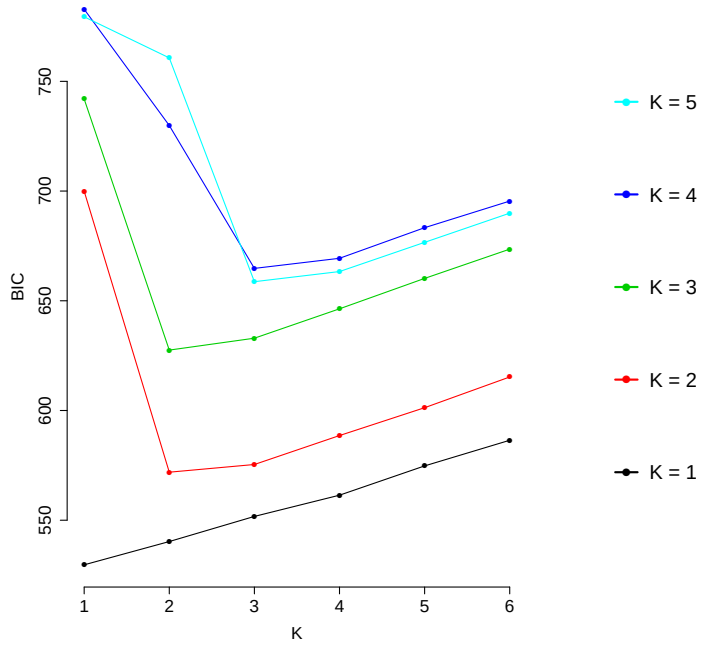


Figure 2.14: The values of BIC for different values of the number of intervals of the true coefficient function.

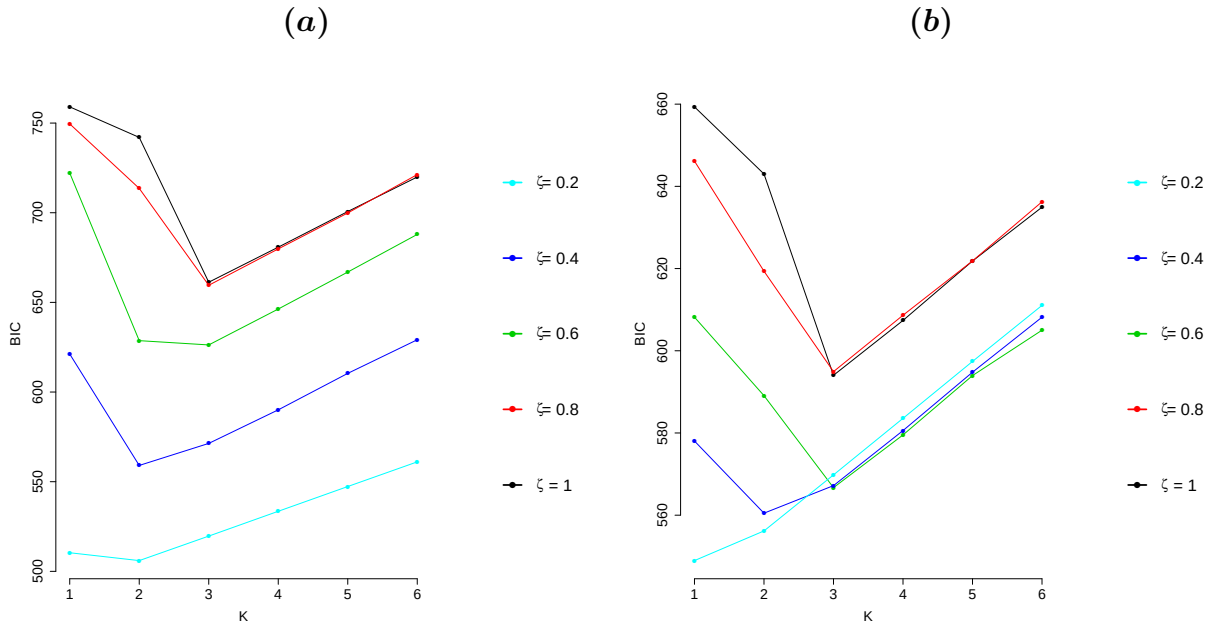


Figure 2.15: The values of BIC for different levels of autocorrelation ζ. *The left plot corresponds to datasets simulated for which the coefficient function is a step function with three intervals (see Figure 2.13). The right plot concerns simulated datasets for which the coefficient function is the Smooth function described in Section 3.1.*

Application on the truffle dataset

Figure 2.17 shows the values of BIC for the truffle dataset described in Section 4.

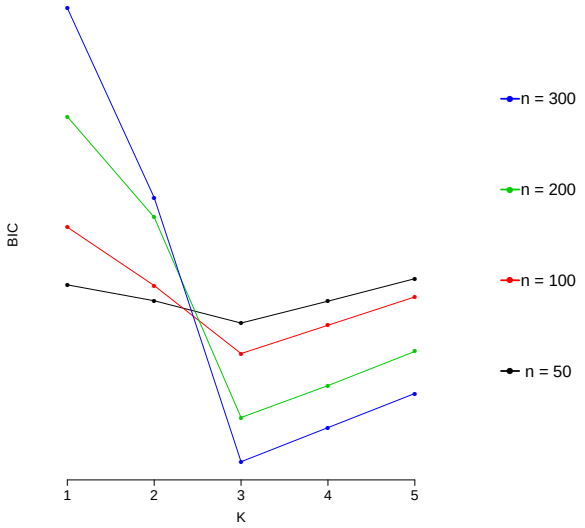


Figure 2.16: The values of BIC for different values of n.

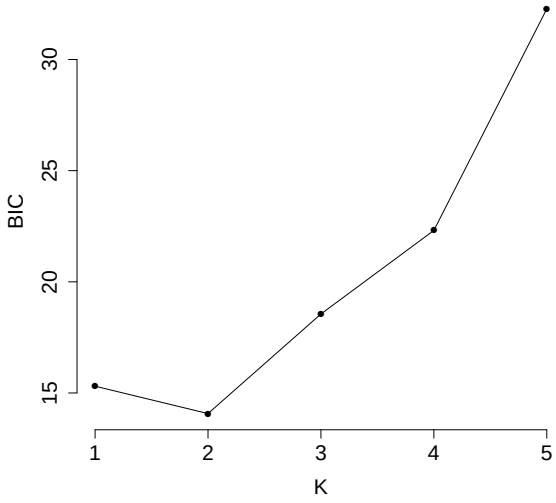


Figure 2.17: The values of BIC for the truffle dataset described in Section 4.

III

Construction d'une loi *a priori* informative

Contents

3.1 Introduction 53
3.2 *A priori* basé sur des pseudo-données 58
3.3 *A priori* basé sur une pénalisation 66
3.4 Implémentation 71
3.5 Résultats numériques 72
3.6 Discussion 86
3.7 Annexe 90

3.1 Introduction

Un point central des approches bayésiennes est le choix d'une distribution *a priori*. Dans de nombreux travaux, cette distribution est choisie au regard de considérations mathématiques. Par exemple, on peut la choisir pour que les lois *a priori* et *a posteriori* soient dans la même famille de lois. D'autres objectifs existent et pour n'en citer que quelques-uns, on mentionne les suivants : minimiser l'information *a priori*, régulariser l'estimateur, obtenir un estimateur plus robuste, encourager l'estimateur à être parcimonieux. Pour plus de détails sur les exemples précédents, on peut consulter Robert (1992).

Dans le cadre d'une approche bayésienne, la loi *a priori* devrait en théorie être déterminée au regard de considérations (non mathématiques) concernant les valeurs plausibles du paramètre du modèle. En effet, la distribution *a priori* est la loi du paramètre indépendamment des données si bien qu'elle est considérée comme représentant une connaissance concernant le paramètre, avant d'avoir observé les données. Le fait de pouvoir prendre en

compte cette connaissance est une importante justification du paradigme bayésien, comme une alternative à l'approche fréquentiste. L'inférence sera alors subjective par opposition à une approche dite objective pour laquelle seuls les données et des choix mathématiques déterminent la loi *a posteriori*.

La prise en compte de connaissances se justifie en particulier lorsque l'inférence statistique se fait sur la base d'un nombre limité de données. Quand les données sont difficiles à obtenir, chères ou rares, il est avantageux de les compléter avec des informations supplémentaires. Ainsi, si le contexte le permet, il est possible de déterminer la loi *a priori* à partir de données historiques ou de résultats d'études similaires précédentes, voir Ibrahim and Chen (2000) ou Chen and Dey (2003) pour un exemple. Lorsque ce n'est pas possible, une autre solution est de réaliser une inférence bayésienne, dite subjective, en prenant en compte des informations d'experts pour construire la loi *a priori*. La distribution *a posteriori* réalise alors un compromis entre l'information apportée par les données et celle apportée par les experts. Cela revient à considérer que la connaissance d'experts est légitime pour décrire et comprendre un phénomène comme complément aux observations. Ceci peut être difficile à accepter et peut être sujet à débat, par exemple Krinitzsky (1993) et Anand (1985) expriment des réticences alors que Meyer and Booker (2001) soulignent l'intérêt de ce type d'approches. Quoiqu'il en soit, de nombreux travaux ont été menés afin de prendre en compte les connaissances d'experts dans des études statistiques. O'Hagan et al. (2006) référencent quelques applications dans des sciences telles que l'économie, l'écologie, la génétique, le renseignement militaire ou encore la gestion du risque (notamment en ce qui concerne l'énergie nucléaire). En écologie, des connaissances d'experts sont utilisées pour des études concernant la présence/absence, en certains lieux, d'espèces en voie de disparition (voir Denham and Mengersen, 2007; O'Leary et al., 2009). L'observation d'individus est alors très rare et il est difficile de bien estimer la localisation et la dynamique de la population. Pour ce genre d'études, les experts ont un avis pertinent concernant la question puisqu'ils connaissent les habitudes et préférences de l'espèce, la topologie du terrain favorisant ou non leur présence, et ont déjà observé des individus en dehors d'un cadre expérimental. Un autre exemple en génétique illustre l'intérêt de ce type informations, lorsque l'objectif est de déterminer les gènes liés à certaines pathologies. Dans ce cas, étant donné le grand nombre de gènes candidats et les corrélations dans les données, il n'est pas évident de sélectionner les bons gènes avec une grande fiabilité. Pour renforcer l'inférence statistique, il a été proposé, entre autres par Pan et al. (2010); Rockova and Lesaffre (2014); Stingo and Vannucci (2010), de prendre en compte dans la distribution *a priori* des réseaux de gènes déterminés par différentes études. De très nombreux exemples de ce type existent dans différents domaines d'application et l'étude de la production de truffes noires du Périgord du chapitre 2 est aussi un cadre où la prise en compte de connaissances d'experts s'avère pertinente.

Pour pouvoir prendre en compte la connaissance d'experts dans un modèle statistique bayésien, il est nécessaire de construire une loi *a priori* reflétant cette connaissance, ce qui nécessite deux grandes étapes de travail. La première consiste à récolter et formaliser l'expertise, on dira qu'il s'agit d'éliciter les experts. Éliciter signifiant "sortir de, mettre en valeur", mais sera ici compris comme le fait de suivre une procédure en collaboration avec un expert afin qu'il puisse partager leurs connaissances. La seconde étape consiste à modéliser l'avis des experts élicités, *i.e.* à déterminer une loi *a priori* reflétant leurs avis. Pour une compréhension de ces deux étapes, on pourra consulter des synthèses de la littérature comme Garthwaite et al. (2005a); Ouchi (2004); Kynn (2005); Jenkinson

(2005), pour n'en citer que quelques-uns.

La littérature sur l'élicitation se décompose en plusieurs thématiques dont une importante est de savoir comment interagir efficacement avec les experts afin d'extraire leur avis. Par exemple en psychologie, l'intérêt est porté sur l'étude de certains biais cognitifs et comportementaux. Pour illustrer, prenons l'exemple d'une enquête basée sur un questionnaire où les répondants doivent renseigner des probabilités, si une question comporte une valeur précise, les répondants auront tendance à répondre en se positionnant par rapport à cette valeur. Les réponses obtenues pourront être globalement assez différentes de celles obtenues si aucune valeur n'avait été suggérée dans le questionnaire (voir Tversky and Kahneman, 1974). Pour en savoir plus sur ces biais, on pourra consulter Kynn (2008). On peut trouver dans la littérature des guides donnant des étapes à suivre pour établir des procédures d'élicitation en évitant certains de ces biais, voir par exemple Winkler et al. (1992); Cooke and Goosens (2000); Low-Choy et al. (2009). Appréhender ces difficultés fait émerger un certain nombre de préceptes dont un des principaux est la simplicité pour les experts, comme défendu par Kadane and Wolfson (1998) :

"The goal of elicitation, as we see it, is to make it as easy as possible for subject-matter experts to tell us what they believe, in probabilistic terms, while reducing how much they need to know about probability theory to do so."

Les développements que nous proposons dans ce chapitre sont principalement guidés par cette considération. L'objectif est de proposer une collaboration avec les experts qui soit naturelle et la plus facile possible pour eux.

Une autre grande partie de la littérature autour de l'élicitation s'intéresse à des développements méthodologiques. Un problème majeur très souvent abordé est de savoir comment agréger l'avis de plusieurs experts. Plusieurs types de modélisation ont été envisagés et on pourra consulter Ouchi (2004) pour une synthèse structurée de cette problématique. Une première manière d'agréger ces informations est l'approche comportementale qui consiste à interagir avec les experts afin qu'ils tendent eux-même vers un consensus, voir par exemple les méthodes *Delphi* (voir Dalkey and Helmer, 1963; Chu and Hwang, 2008) et *Nominal Group* (Delbecq and Van de Ven, 1971). La seconde manière consiste à agréger les informations élicitées mathématiquement dans le modèle. Par exemple, Genest and Zidek (1986) proposent de modéliser conjointement les informations d'experts, alors que d'autres comme Burgman et al. (2011) proposent de prendre en compte une information moyennée des experts. Cependant, d'autres types d'approches existent, voir par exemple Hunns and Daniels (1981) et Cooke (1991) pour plus de détails.

Un autre angle sous lequel appréhender cette littérature consiste à faire une distinction entre les approches dites directes et indirectes. Dans certains cas (rares en pratique), l'expert est considéré comme un statisticien ou très familier avec des concepts de statistiques et de probabilités. Dans un tel cas, il est concevable que l'expert soit à même de renseigner directement la valeur d'un hyperparamètre (Zellner, 1972). Une telle approche est dite directe dans le sens où l'expert renseigne lui-même directement un paramètre du modèle (voir par exemple Winkler, 1967; Fleishman et al., 2001; Kadane et al., 1980; O'Leary et al., 2008). Ce type d'approches est efficace puisque l'expert est capable de formuler fidèlement sa connaissance en termes mathématiques et la modélisation n'a pas à être sensiblement modifiée par rapport à un modèle sans connaissance d'experts. Cependant,

cette situation n'est pas commune et il est souvent nécessaire de devoir élaborer des méthodes plus complexes pour obtenir une formulation mathématique de l'avis des experts. Ainsi, par opposition aux approches directes, les approches dites indirectes consistent à récolter des informations qui ne sont pas directement des valeurs d'hyperparamètres, ni même une caractéristique de la distribution *a priori*. Un travail non trivial de "traduction" est alors nécessaire pour définir la distribution *a priori* à partir de ces informations. Contrairement à une approche directe, ce type d'approche permet généralement d'avoir un échange plus efficace avec l'expert. En effet, l'expert est généralement peu familier des notions statistiques et probabilistes, si bien qu'il ne lui est pas évident de renseigner des valeurs d'hyperparamètres qui reflètent réellement ce qu'il pense. Au lieu de questionner l'expert sur des quantités théoriques (approche directe), il est plus approprié de le questionner sur des quantités observables (approche indirecte) avec lesquelles il est plus familier. Par exemple, avec des questions simples (Huber, 1974), on peut extraire de l'expert ce qui correspond pour lui aux quantiles de la loi des données (voir O'Hagan et al., 2006; Albert et al., 2012). Cela permet d'approcher ce que serait la loi prédictive des données pour l'expert.

Dans le cadre du modèle de régression linéaire, de nombreuses procédures d'élicitation ont été développées. Certaines cherchent à prendre en compte des objectifs particuliers comme la sélection de variables (Garthwaite and Dickey, 1992). D'autres procédures sont établies pour éliciter des connaissances concernant un paramètre, comme la pente ou le paramètre de variance (Kadane et al., 1980; Garthwaite and Dickey, 1991). Concernant le modèle de régression linéaire multiple, James et al. (2010) souligne la difficulté d'une procédure d'élicitation directe. En effet, quand le plan d'expérience n'est pas orthogonal, il n'est pas évident d'interpréter les coefficients de pente. Or les méthodes directes sont basées sur la capacité des experts à interpréter les coefficients du modèle. Il est tout de même possible de contourner ce problème et d'établir des procédures d'élicitation directe dans le cadre de ce modèle. Par exemple, au lieu de demander directement la valeur d'un hyperparamètre, il est possible de demander aux experts de renseigner un avis concernant certaines caractéristiques des coefficients (voir Kuhnert et al., 2005; Martin et al., 2005; O'Leary et al., 2008). En particulier, le signe des coefficients est plus simple à interpréter et semble moins sensible à la structure du *design*. Dans le cadre de ce modèle, Kadane and Wolfson (1998) souligne l'intérêt d'une approche indirecte et on trouvera plus de détails concernant ce type d'approches dans Carlin and Louis (2000) ou O'Hagan et al. (2006).

Dans ce chapitre, nous nous placerons dans le cadre du modèle Bliss (détaillé dans le chapitre 2) et discuterons du problème de la détermination de la loi *a priori* afin de prendre en compte la connaissance d'experts. Pour rappel, le modèle Bliss est un cas particulier du modèle de régression linéaire pour des données fonctionnelles. L'objectif principal de ce modèle est de fournir des estimations simples afin de dégager clairement les périodes sur lesquelles des covariables fonctionnelles $x(\cdot)$ ont un impact majeur sur une variable réelle y. Supposons disposer d'observations y_i de la variable réponse et d'observations $x_i(t)$ d'une covariable fonctionnelle pour certaines valeurs de $t \in [0,1]$ et pour $i = 1, \ldots, n$. Le modèle Bliss est basé sur une décomposition adaptative de la fonction coefficient $\beta(\cdot)$ du modèle de régression linéaire donné dans le chapitre 2 par (2.3). Cette décomposition contraint $\beta(\cdot)$ à être une fonction constante par morceaux avec K intervalles :

$$\beta(t) = \sum_{k=1}^{K} b_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}_{\mathcal{I}_k}(t),$$

où les b_k sont des réels et chaque $\mathcal{I}_k$ est un intervalle de $[0,1]$. Le modèle Bliss est donné par

$$y_i|x_i(\cdot),\mu,b,\sigma^2,\mathcal{I} \overset{ind}{\sim} \mathcal{N}\left(\mu + x_i(\mathcal{I})^T b\ ,\ \sigma^2\right),\ \text{for } i=1,\dots,n, \tag{3.1}$$

où $x_i(\mathcal{I})$ est un vecteur dont la k^e coordonnée est $\frac{1}{|\mathcal{I}_k|}\int_{\mathcal{I}_k} x_i(t)\mathrm{d}t$. Comme dans le chapitre 2, les intervalles $\mathcal{I}_k$ sont ici caractérisés par deux paramètres, un milieu m_k et une demi-longueur ℓ_k :

$$\mathcal{I}_k = [m_k - \ell_k, m_k + \ell_k].$$

La distribution *a priori* des paramètres du modèle utilisée dans le chapitre 2 est donnée de manière hiérarchique par

$$\begin{aligned}
\mu|\sigma^2 &\sim \mathcal{N}\left(0, v_0\sigma^2\right),\\
b|\sigma^2,\mathcal{I} &\sim \mathcal{N}_K\left(0, \sigma^2\Sigma(\mathcal{I})\right),\\
\pi(\sigma^2) &\propto 1/\sigma^2\\
m_k &\overset{i.i.d.}{\sim} \mathrm{Unif}\left(\mathcal{T}\right), \quad k=1,\dots,K,\\
\ell_k &\overset{i.i.d.}{\sim} \mathrm{Exp}(a\times|\mathcal{T}|), \quad k=1,\dots,K,
\end{aligned} \tag{3.2}$$

où $\Sigma(\mathcal{I})$ est la matrice de covariance inspirée à l'*a priori* de *Ridge Zellner*. Cette distribution a été déterminée afin d'obtenir une loi *a priori* faiblement informative. Ainsi, la distribution *a posteriori*

$$\begin{aligned}
\pi(\mu,b,\sigma^2,m,\ell|y) \propto &\left(\sigma^2\right)^{-\frac{1}{2}(n+K+1)-1} |\Sigma(\mathcal{I})|^{-1/2}\\
&\times \exp\left\{-\frac{1}{2\sigma^2}\left[\mathrm{SCR} + \mu^2 v_0^{-1} + b^T\Sigma(\mathcal{I})^{-1}b\right]\right\}\pi(\ell)
\end{aligned} \tag{3.3}$$

est principalement déterminée par l'information apportée par les observations, où $SCR = \sum_{i=1}^n (y_i - \mu - x_i(\mathcal{I})^T b)^2$.

Dans le chapitre précédent, ce modèle est utilisé pour étudier l'impact des précipitations sur la production de truffes noires du Périgord (données fournies par J. Demerson). Pour cette étude, peu de données sont disponibles et l'estimation est donc un problème statistique complexe. En effet, seulement 13 années sont observées pour estimer 11 paramètres (lorsqu'on fixe $K=3$ intervalles). D'où l'intérêt d'utiliser une procédure d'élicitation pour pallier ce manque de données. De plus, les experts ont des connaissances biologiques concernant la croissance de la truffe et ses mécanismes complexes de reproduction. Ces connaissances sont clairement utiles pour comprendre la sensibilité de la truffe à l'humidité et sont donc importantes à prendre en compte dans cette étude.

L'objectif de ce chapitre sera d'établir des procédures d'élicitation en respectant deux critères importants. Le premier est d'avoir une interaction avec les experts aussi simple que possible pour eux. Le second point est de contrôler l'impact sur l'inférence de la prise en compte de connaissances *a priori*. Dans ce qui suit, deux approches sont proposées pour éliciter les experts en respectant ces objectifs. La première, détaillée en section 3.2, est basée sur l'idée qu'une procédure simple pour les experts consiste à les questionner sur des quantités observables, et non sur des paramètres du modèle. Les experts donnent leur avis en imaginant des données, dites pseudo-données. Nous proposons de prendre en compte dans un même modèle, ces pseudo-données renseignées et les données observées. Pour cela,

nous proposons en section 3.2.1 un modèle faisant intervenir une vraisemblance pondérée de telle sorte que les pseudo-données aient moins d'importance que les données observées dans l'inférence. Une écriture équivalente de cette approche permet de constater que cela revient à considérer comme loi *a priori* sur θ, la loi *a posteriori* de θ sachant les pseudo-données, pondérée par un système de poids. Nous justifions en section 3.2.2 ce choix en donnant des propriétés de la distribution *a posteriori*, qui paraissent importantes lorsqu'on souhaite relativiser l'impact de la prise en compte des pseudo-données sur l'inférence. Nous détaillons en section 3.2.3 comment nous proposons de calibrer les poids, de sorte qu'ils vérifient certaines propriétés qui nous semblent importantes dans notre contexte d'application. Pour la seconde approche, détaillée en section 3.3, nous demandons l'avis des experts sur certaines caractéristiques de la fonction coefficient : son support et son signe sur certaines périodes. Nous proposons une nouvelle distribution *a priori* pour prendre en compte ce type d'avis. Cette distribution fait intervenir un terme qui apparaît comme un terme de pénalisation dans la distribution *a posteriori*. Les fonctions coefficient qui ne correspondent pas à l'avis des experts sont pénalisées. Nous détaillons en section 3.3.1 et en section 3.3.2 comment nous choisissons ce terme de pénalisation. En particulier, deux approches sont proposées pour calibrer le paramètre correspondant à l'intensité de la pénalisation. La première approche consiste à le calibrer en utilisant une validation croisée bayésienne. Pour la seconde, nous proposons d'introduire une loi *a priori* pour ce paramètre. En section 3.4, nous décrivons comment les méthodes proposées en section 3.2 et en section 3.3 sont implémentées. Nous appliquons ces approches sur des données simulées en section 3.5.1 afin d'illustrer l'impact sur l'inférence de la prise en compte de l'avis d'experts. En section 3.5.2, ces méthodes sont appliquées sur les données de trufficulture afin de déterminer l'impact des précipitations sur la production de truffes en s'aidant des connaissances des experts. Nous discuterons finalement des deux approches et leurs performances en section 3.6.

3.2 *A priori* basé sur des pseudo-données

Dans le cas du modèle de régression linéaire sur données fonctionnelles, comme dans celui du modèle Bliss (3.1), l'interprétation de la fonction coefficient $\beta(\cdot)$ n'est pas triviale. Il n'est donc pas évident pour un expert d'avoir un avis *a priori* sur les valeurs de $\beta(t)$ pour $t \in [0, 1]$. Ainsi, une approche directe semble complexe à mettre en œuvre dans notre contexte. Cependant, au-delà de l'interprétation de ce coefficient fonctionnel, les experts peuvent néanmoins avoir un avis concernant l'impact de la covariable fonctionnelle sur la variable réponse. L'élicitation a alors pour but d'obtenir le maximum d'informations concernant cet avis.

Comme exprimé précédemment, une manière efficace de questionner un expert est de lui poser des questions en termes de quantités observables, avec lesquelles il sera à l'aise pour raisonner. Une des approches les plus utilisées consiste à extraire la connaissance de l'expert concernant la distribution des données en le questionnant (indirectement) sur les quantiles de cette distribution. Dans ce cas, l'inconvénient réside dans la difficulté pour l'expert à avoir une idée des quantiles d'une loi. En théorie, les questions posées à l'expert sont intuitives pour lui et devraient permettre d'obtenir de bons résultats (voir Kynn, 2005). Cependant, l'esprit humain a tendance à mal appréhender les probabilités proches

de 0 ou de 1 (voir Kahneman and Tversky, 1979; Alpert and Raiffa, 1982). Ainsi, ce type d'approches peut avoir un biais si des quantiles extrêmes sont élicités. À l'opposé, si des quantiles sont élicités uniquement autour du *centre* de la distribution, les queues de la distribution peuvent être mal estimées. De plus, dans certains cas, plusieurs distributions usuelles peuvent correspondre à un ensemble de quantiles élicités. De ces distributions, on obtient des distributions *a posteriori* différentes, ce qui pose un problème conceptuel (voir Berger, 1985).
C'est pourquoi, pour extraire l'avis d'expert, nous proposons une approche basée sur des données imaginées par les experts, dites pseudo-données. L'idée est la suivante : pour une valeur fixée de la covariable x, l'expert va faire appel à ce qu'il sait du phénomène sous-jacent pour *prédire* une valeur de y. On s'attend les pseudo-données fournies par l'expert vont nous renseigner sur ses connaissances. Cependant, la valeur que donne l'expert est entachée de plusieurs sources d'incertitudes importantes (O'Hagan et al., 2006). En effet, on pourrait émettre un doute sur la capacité d'un expert à donner une valeur pour y avec une bonne précision. Nous proposons une modélisation qui prend en compte ce problème de telle sorte que les pseudo-données soient modélisées par une loi avec une variance relativement grande.

Par la suite, nous noterons les données observées par y_i et $x_i(\cdot)$, pour $i = 1, \ldots, n$, et les pseudo-données de l'expert e par y_i^e et $x_i^e(\cdot)$, pour $i = 1, \ldots, n_e$ et $e = 1, \ldots, E$. Nous considérons que les pseudo-données sont différentes par nature des données observées car elles sont entachées d'une plus grande incertitude. Nous souhaitons faire ressortir dans notre modélisation que la distribution de $y^e = (y_1^e, \ldots, y_{n_e}^e)$ admet une plus grande variance que la distribution de y, pour chaque expert e. Par rapport à cet objectif, la modélisation que nous proposons est à la croisée de l'approche *Power Prior* (Ibrahim and Chen, 2000) et de la régression par vraisemblance pondérée (Hu and Zidek, 1995).
L'approche *Power Prior* a pour but de construire une loi *a priori* à partir de données historiques D_0 tout en pondérant l'apport des données historiques par rapport aux données observées. Pour un a_0 donné, la distribution *Power Prior* d'un paramètre θ (prise comme distribution *a priori*) est donnée par

$$\pi(\theta|D_0, a_0, c_0) \propto L(\theta|D_0)^{a_0}\, \pi_0(\theta|c_0),$$

où $\pi_0(\theta|c_0)$ est une distribution *a priori* initiale par rapport à un hyperparamètre c_0 et $L(\theta|D)$ est la vraisemblance des données D pour une valeur de θ. Pour $a_0 = 1$, la distribution *Power Prior* est la distribution *a posteriori* de θ sachant les données historiques. Si $a_0 = 0$, $\pi(\theta|D_0, a_0) = \pi_0(\theta|c_0)$ ce qui équivaut à ne pas prendre en compte les données historiques. Pour plus de détails, Ibrahim et al. (2015) proposent une synthèse des résultats et des variantes de cette méthode.
Quant à la régression par vraisemblance pondérée (Hu and Zidek, 1995), elle consiste à faire varier l'influence des données dans l'analyse en s'autorisant à commettre plus ou moins d'erreurs de prédiction sur certaines données. Augmenter l'erreur de prédiction sur certaines données (introduire un biais) a pour objectif de baisser la variance de l'estimateur. L'estimateur est plus robuste et réalise un meilleur compromis biais/variance. Pour des données observées $y = (y_1, \ldots, y_n)$ et un paramètre θ, la vraisemblance pondérée est donnée par

$$L(y|\theta; \lambda_1, \ldots, \lambda_n) = \prod_{i=1}^{n} p(y_i|\theta)^{\lambda_i}$$

où $p(y_i|\theta)$ est la vraisemblance de y_i pour une valeur de θ et les λ_i sont des poids à calibrer (voir Hu and Zidek, 2002). Un des développements importants autour de cette approche, concerne les résultats asymptotiques de l'estimateur de maximum de vraisemblance (Wang et al., 2004).

3.2.1 Modèle

Dans cette section, nous décrivons comment nous modélisons les données observées et les pseudo-données. Soient w_i^e des poids associés à chacune des pseudo-données y_i^e. La calibration de ces poids w_i^e est discutée en section 3.2.3. Rappelons que $\theta = (\mu, b, \sigma^2, m, \ell)$. Les données observées sont modélisées par le modèle Bliss (3.1). De plus, nous considérons que la connaissance de chaque expert peut être approchée par ce même modèle (3.1). Ainsi, on suppose que la distribution des pseudo-réponses de l'expert e est donnée par

$$y_i^e|x_i^e(\cdot)\mu, b, \sigma^2, \mathcal{I} \overset{ind}{\sim} \mathcal{N}\left(\mu + x_i^e(\mathcal{I})^T b, \sigma^2\right), \text{ for } i = 1, \ldots, n_e. \tag{3.4}$$

La distribution *a priori* de θ n'est pas changée et est donnée par (3.2). L'idée principale de cette approche est de définir la vraisemblance jointe des données observées et des pseudo-données comme une vraisemblance pondérée :

$$L(y, y^1 \ldots, y^E|\theta; w) = \prod_{i=1}^{n} p(y_i|\theta) \times \prod_{e=1}^{E} \prod_{i=1}^{n_e} p(y_i^e|\theta)^{w_i^e}, \tag{3.5}$$

où $w = (w^1, \ldots, w^E)$ avec $w^e = (w_1^e, \ldots w_{n_e}^e)$ pour $e = 1, \ldots, E$. Une formulation équivalente de cette approche est de considérer le modèle pour lequel la distribution des données observées est donnée par (3.1) et la distribution *a priori* de θ est donnée par

$$\pi(\theta|y^1, \ldots, y^E; w) = \pi_0(\theta) \times \prod_{e=1}^{E} \prod_{i=1}^{n_e} p(y_i^e|\theta)^{w_i^e}, \tag{3.6}$$

où $\pi_0(\cdot)$ est la densité de la distribution *a priori* (3.2). Selon cette vision, on choisit comme distribution *a priori*, la distribution *a posteriori* de θ sachant les pseudo-données, ce qui correspond à une approche bayésienne séquentielle (voir Berger, 1985, chapitre 7). Cette formulation permet de voir clairement le lien avec l'approche *Power Prior* pour laquelle la distribution *a priori* contient l'information d'observations d'une étude antérieure. La similitude est d'autant plus claire lorsqu'on considère l'approche *Power Prior* pour L_0 jeux de données historiques (Ibrahim and Chen, 2000) :

$$\pi(\theta|D_0, a_0, c_0) \propto \prod_{k=1}^{L_0} L(\theta|D_{0k})^{a_{0k}} \pi_0(\theta|c_0), \tag{3.7}$$

où $D_0 = (D_{01}, \ldots, D_{0L_0})$ est l'ensemble des jeux de données historiques et $a_0 = (a_{01}, \ldots, a_{0L_0})$ est un vecteur de poids à déterminer. Plusieurs approches ont été envisagées pour calibrer a_0. Dans certains travaux, il a été proposé d'attribuer une loi *a priori* sur a_0 (voir entre autres Ibrahim and Chen, 2000). D'autres auteurs ont proposé de fixer a_0 (De Santis, 2006). Pour le fixer, Chen et al. (2006) s'inspirent d'un lien entre la méthode *Power Prior* et une approche introduisant les données historiques de manière hiérarchique.

Une spécificité du modèle que nous proposons par rapport à (3.7) est qu'un poids est associé à chacune des pseudo-données et non à un ensemble de données. Par rapport à la méthode par vraisemblance pondérée, nous ne pondérons pas les données observées mais des données supplémentaires.
De plus, notre approche se distingue des travaux de (Ibrahim and Chen, 2000) et (Hu and Zidek, 1995) puisque nous travaillons avec des pseudo-données. Nous ne considérons pas des pseudo-données comme des données observées ou des données historiques. Ainsi, la détermination des poids w_i^e doit prendre en compte cette différence. Avant de présenter notre manière de calibrer les poids, nous introduisons deux propriétés importantes.

3.2.2 Propriétés *a posteriori*

Le choix de cette modélisation est fait au regard de deux propriétés détaillées dans cette section. Tout d'abord, posons $w^e = n_e^{-1} \sum_{i=1}^{n_e} w_i^e$, la moyenne des poids de l'expert e et W^e la matrice diagonale dont le i^e élément de la diagonale est w_i^e. La première propriété est en lien avec la distribution *a posteriori* de θ (sachant les données observées et les pseudo-données) qui est proportionnelle à

$$\begin{aligned}\pi(\theta|y, y^1, \ldots, y^E; w) &\propto \prod_{i=1}^{n} p(y_i|\theta) \times \prod_{e=1}^{E} \prod_{i=1}^{n_e} p(y_i^e|\theta)^{w_i^e} \times \pi_0(\theta) \\ &\propto \left(\sigma^2\right)^{-\frac{1}{2}(n+\sum_{e=1}^{E} n_e w^e + K + 1) - 1} |\Sigma(\mathcal{I})|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[\mathrm{SCR} + \sum_{e=1}^{E} \mathrm{SCR}_e + \mu^2 v_0^{-1} + b^T \Sigma^{-1} b\right]\right\} \pi(\ell)\end{aligned} \tag{3.8}$$

où $\mathrm{SCR} = \sum_{i=1}^{n} (y_i - \mu - x_i(\mathcal{I})^T b)^2$, $\mathrm{SCR}_e = \sum_{i=1}^{n_e} w_i^e (y_i^e - \mu - x_i^e(\mathcal{I})^T b)^2$ et w^e est la moyenne des w_i^e, pour $i = 1, \ldots, n_e$. On peut comprendre σ^2 / w_i^e comme la variance de l'erreur pour une pseudo-donnée. Plus la variance est grande par rapport à σ^2, moins la pseudo-donnée compte.

Propriété 3.1. *Comme $\sum_{e=1}^{E} n_e w^e$ joue le même rôle que n dans la distribution a posteriori* (3.8)*, ce terme s'interprète comme un équivalent échantillon. En ce sens, si $w_i^e = 1$ pour une pseudo-donnée alors elle aura autant de poids qu'une donnée observée.*
A l'opposé, si les poids w_i^e sont tous nuls, la distribution a posteriori de θ ne dépend donc pas des pseudo-données et est proportionnelle à

$$\pi(\theta|y, y^1, \ldots, y^E; w) \propto \pi(\theta|y)$$

où $\pi(\theta|y)$ est donnée par (3.3)*.*

Pour la seconde propriété, considérons pour simplifier que les données sont centrées par rapport à leurs poids respectifs. L'espérance *a posteriori* de b s'écrit alors comme

$$\mathbb{E}\left(b|y, y^1, \ldots, y^E, \mathcal{I}\right) = \hat{b}_1 + \sum_{e=1}^{E} \hat{b}_{2,e}, \tag{3.9}$$

où

$$\begin{aligned}\hat{b}_1 &= M_w^{-1} x(\mathcal{I})^T y \\ \hat{b}_{2,e} &= M_w^{-1} x^e(\mathcal{I})^T W^e y^e \\ M_w &= \Sigma(\mathcal{I})^{-1} + x(\mathcal{I})^T x(\mathcal{I}) + \sum_{e=1}^{E} x^e(\mathcal{I})^T W^e x^e(\mathcal{I}).\end{aligned}$$

Cette expression nous permet de constater la propriété suivante.

Propriété 3.2. *Conditionnellement aux intervalles, l'estimateur bayésien de b se décompose comme une partie apportée par les réponses observées au travers de* $\hat{b}_1$ *et une partie apportée par les pseudo-réponses au travers de* $\hat{b}_{2,1}, \dots \hat{b}_{2,E}$*. De plus, l'apport des pseudo-données est pondéré par les poids* w_i^e *au travers des matrices* W^e*. Dans le cas particulier où la matrice* W^e *est la matrice nulle, le poids de chacune des pseudo-données est nul et l'espérance* (3.9) *coïncide avec l'expression standard de l'estimateur bayésien du coefficient de pente d'une régression linéaire (voir, par exemple, Marin and Robert, 2007).*

3.2.3 Calibration des poids

Étant donné les propriétés précédentes, on constate que les poids w_i^e ont une importance majeure dans la distribution *a posteriori*. Un point important de l'approche proposée est donc de déterminer ces poids.
Pour ce qui est de la calibration du poids a_0 pour l'approche *Power Prior*, Ibrahim et al. (2015) discutent de différentes possibilités comme considérer une loi *a priori* sur le a_0 ou sélectionner a_0 au regard d'un critère de vraisemblance pénalisé. Concernant la régression par vraisemblance pondérée, les poids peuvent par exemple être calibrés afin d'obtenir des estimateurs robustes (voir, parmi d'autres, Markatou et al., 1998). Dans notre contexte, nous avons deux spécificités importantes. La première est que nous ne traitons pas ici des données observées ou historiques, mais des pseudo-données. Ces pseudo-données sont, par nature, différentes des précédentes, ce qui doit induire une différence dans notre manière de calibrer les poids. Une autre différence par rapport au contexte de l'approche *Power Prior* est que nous avons besoin de calibrer un poids par pseudo-données.

Dans le cadre d'une procédure d'élicitation, plus proche de notre contexte, plusieurs manières pour calibrer des poids ont été envisagées (voir Cooke, 1991), et Winkler (1968) en cite quatre :

1. Donner aux experts des poids égaux.
2. Classer les experts par *préférence* et associer des poids en conséquence.
3. Laisser les experts s'attribuer des poids.
4. Établir une règle de notation (*scoring rule*).

Le premier point revient à refuser d'attribuer des poids aux experts.
Concernant le deuxième point, sans considération supplémentaire, il ne paraît pas direct

de passer d'une hiérarchie à des poids. Pour compléter cette idée, on pourrait penser à juger de la *qualité* d'un expert par rapport aux données observées. Si les pseudo-données d'un expert ressemblent aux données observées, on pourrait associer un poids élevé à ses données. Cependant ce type d'approches présente certains inconvénients. Premièrement, cela revient à utiliser plusieurs fois les données, ce qui est en théorie proscrit dans une inférence bayésienne. Ensuite, l'avis des experts est ainsi considéré moins pertinent s'il n'est pas en accord avec l'observation, ce qui revient à minimiser l'avis des experts.
La troisième idée semble rencontrer certains biais comportementaux décrits dans Kynn (2008). Si un expert s'attribue des poids tout seul sans contrôle, il peut être amené à placer une trop grande confiance en lui. Pour illustrer ce point, on pourra consulter De Groot (1974) qui propose une procédure itérative où les experts se notent mutuellement et révisent leurs notes aux regards des notes des autres experts. L'auteur établit alors les notes limites que les experts s'attribueraient en répétant le processus indéfiniment.
Concernant le quatrième point, une règle de notation se définit comme le fait d'attribuer un score pour juger de la *qualité* des réponses des experts (voir Murphy and Winkler, 1970; Lindley, 1982; Gneiting and Raftery, 2007). Pour simplifier, prenons l'exemple d'une situation où nous demandons aux experts de renseigner la distribution d'une variable aléatoire pouvant prendre m valeurs. Notons p_i les probabilités fournies par l'expert, pour $i = 1, \dots, m$. Un score $R(p, i)$ est construit à partir des informations données par l'expert et des observations (ce qui n'est pas conforme à la théorie bayésienne). Supposons que l'expert pense à une distribution q mais qu'il n'est capable que de renseigner la distribution p, une approximation de q. Une heuristique pour établir R, est que $\mathbb{E}_q\, R(p, i)$ soit maximum (en p) lorsque $p = q$ (Cooke, 1991). La règle de notation R vient alors pénaliser les distributions p qui sont trop *éloignées* de q. L'intérêt est de *récompenser* ou de *pénaliser* l'expert en fonction de son score afin de l'encourager à être honnête et à renseigner des réponses qui soient en accord avec son intuition (Garthwaite et al., 2005b). Une synthèse récente de la littérature concernant le *scoring* est donnée dans Carvalho (2016).

Dans ce qui suit, nous cherchons à déterminer des poids pour chacune des pseudo-données fournies par les experts. Il est important de noter que les poids à déterminer concernent des pseudo-données alors que les poids sont généralement attribués aux experts. Le fait d'avoir un poids par pseudo-donnée permet une certaine finesse dans notre modélisation. Nous nous inspirons du point 3 et nous proposons de déterminer les poids en nous basant sur la *confiance* des experts en leurs avis. De plus, les poids seront déterminés en prenant en compte deux considérations supplémentaires, détaillées ci-dessous. La première concerne la dépendance entre les experts et la seconde concerne les poids relatifs des pseudo-données par rapport à celui des données observées.

Certitude des experts Pour la détermination des poids, nous souhaitons ici prendre en compte la confiance qu'ont les experts en leurs pseudo-données. Au lieu du terme "confiance" qui est associé à une notion spécifique en statistique, nous préférerons le terme "certitude". L'objectif est ici d'extraire de chaque expert e, sa certitude concernant sa i^e pseudo-donnée, qu'on notera c_i^e. Pour interagir avec les experts, nous nous basons sur les propriétés *a posteriori* décrites en section 3.2.2 pour les aider à avoir une interprétation intuitive de ces certitudes. En première approche, nous considérons avec les experts que les poids des pseudo-données dans l'analyse correspondent à leurs certitudes ($w_i^e = c_i^e$).

Dans ce cas, les experts peuvent interpréter une certitude comme suit :

- Si $c_i^e = 0$, la pseudo-donnée associée ne comptera pas dans l'analyse. Cette valeur est à associer à un scénario invraisemblable, ce qui ne représente pas d'intérêt ici pour l'expert, mais lui permet d'appréhender le sens d'une certitude faible.
- Si $c_i^e = 1$, la pseudo-donnée associée comptera autant qu'une donnée observée. Si un expert pense à un scénario de précipitations et de production qu'il a observé dans le passé, il lui est suggéré de choisir une telle valeur de certitude.
- Si $c_i^e = 1/2$, la pseudo-donnée associée comptera *conceptuellement* comme une demie-donnée. Si un expert donne deux données avec cette certitude, la paire de pseudo-données comptera comme une donnée observée, au sens de l'erreur d'ajustement $(y_i^e - \mu - x_i^e(\mathcal{I})^T b)^2$. En effet, supposons prendre en compte deux nouvelles pseudo-données avec des certitudes de $1/2$ et que pour chacune d'elles on fasse une erreur de k avec le modèle. Dans ce cas, on augmente le terme exponentiel dans (3.8) de k, en augmentant la valeur de SCR_e. La valeur du terme exponentiel serait augmentée de la même manière au travers de SCR en prenant en compte une nouvelle observation pour laquelle une erreur de k aurait été faite. Si un expert renseigne n_e pseudo-données et qu'il assigne à chacune d'elle une certitude $1/n_e$, son jeu de pseudo-données comptera comme une donnée supplémentaire. Cette dernière interprétation sera plus intuitive pour un statisticien qui pourra interpréter n_e comme un équivalent échantillon.

Étant donné les biais comportementaux, il paraît compliqué de demander aux experts de choisir eux-mêmes leurs poids dans l'analyse. On propose donc d'affiner notre approche en prenant en compte les deux considérations suivantes.

Dépendance entre experts La première considération concerne la dépendance entre les experts (voir French (1985)) qui semble importante à prendre en compte dans notre contexte. Si les experts travaillent dans une même équipe ou s'inscrivent dans un même mouvement de pensée, leurs avis auront tendance à être similaires. On peut alors souhaiter prendre en compte cette similitude dans notre manière de modéliser l'avis des experts. En effet, imaginons le cas critique où nous élicitons plusieurs fois le même expert, nous voulons éviter que le poids de son avis soit multiplié afin qu'il ne prenne pas trop d'importance par rapport aux observations. Donc si plusieurs experts sont hautement dépendants, on souhaite pouvoir les considérer comme un seul expert dans le sens où ils partagent un même avis. Ainsi, il peut être souhaitable de corriger leurs poids respectifs afin que leur avis ne soit pas artificiellement dominant par rapport aux observations. Pour prendre en compte cette dépendance, nous proposons de corriger la certitude des experts par le terme correctif multiplicatif

$$\frac{1}{1 + \sum_{j \neq i} r_{i,j}^2}, \tag{3.10}$$

où $r_{i,j} \in [-1, 1]$ est un coefficient de dépendance entre l'expert i et l'expert j. Ainsi, si deux experts i et j dépendants, avec $r_{i,j} = 1$, fournissent chacun une pseudo-donnée et une certitude 1, alors en prenant en compte le terme correctif (3.10), égal à $1/2$ ici, nous considérons que les deux experts n'ont fournis qu'une seule pseudo-donnée. Les coefficients

de dépendance $r_{i,j}$ sont déterminés empiriquement à partir des interactions connues entre les experts.

Exemple 1. *Supposons observer 5 données* (y_i) *et éliciter trois experts fournissant chacun 5 pseudo-données* (y_i^e)*. Supposons que ces experts soient hautement dépendants. La table 3.1 référence les données, les pseudo-données et les certitudes des experts. Dans cet exemple, le poids des données est* 5 $(= n)$*, et considérons les poids des pseudo-données égaux aux certitudes des experts* $(w_i^e = c_i^e)$*. Sans prendre en compte la dépendance entre les experts, leur poids total* $\left(\sum_{e=1}^{5}\sum_{i=1}^{5} w_i^e\right)$ *est* $0.5 \times 5 \times 3 = 7.5$*. On choisit les poids de la manière suivante :*

$$w_i^e = \frac{c_i^e}{1 + \sum_{j \neq i} r_{i,j}^2},$$

et si on fixe $r_{1,2} = r_{1,3} = r_{2,3} = 1$*, alors* $\frac{1}{1+\sum_{j\neq i} r_{i,j}^2} = 1/3$*, d'où les poids* w_i^e *sont égaux dans cet exemple à* 0.167 *et le poids total des experts est* $2.5 (= 0.167 \times 5 \times 3)$*.*

Table 3.1: ***Exemple jouet de données observées et de pseudo-données.*** *Les experts 1 à 3 renseignent des pseudo-données* y_i^e *et des certitudes* c_i^e*.*

Données	Expert 1		Expert 2		Expert 3	
y_i	y_i^1	c_i^1	y_i^2	c_i^2	y_i^3	c_i^3
10.1	15	0.5	15	0.5	14	0.5
8.2	10.5	0.5	11	0.5	10	0.5
9.7	10.5	0.5	11	0.5	11	0.5
16.8	16	0.5	16	0.5	15	0.5
23	19	0.5	20	0.5	20	0.5

Poids relatifs des pseudo-données Dans le paragraphe précédent, nous expliquons que les pseudo-données peuvent prendre un poids important par rapport aux données observées, dans le cas d'experts dépendants et si cette dépendance n'est pas prise en compte. Un phénomène similaire apparaît quand le nombre de pseudo-données devient trop grand face aux données observées. Pour éviter ce problème, nous proposons de nouveau de corriger la certitude des experts par un terme correctif multiplicatif

$$\frac{n}{n_e \times E}, \tag{3.11}$$

où n est le nombre de données observées, E est le nombre d'experts et n_e est le nombre de pseudo-données fournies par l'expert e. Ainsi, si un expert fournit n pseudo-données, le terme correctif (3.11) qui lui est associé sera $1/E$. Le but de ce terme correctif est que le poids des pseudo-données ne dépasse pas le poids des données observées.

Exemple 2. *Dans l'exemple 1, supposons que l'expert 1 (resp. 2 et 3) fournisse 5 (resp. 10 et 15) pseudo-données. Si nous ne prenons plus en compte la dépendance entre experts* $(w_i^e = c_i^e)$*, alors le poids total des experts auraient été de 15. Si nous prenons en compte le terme correctif 3.11, alors*

$$w_i^e = c_i^e \frac{n}{n_e \times E}$$

et le poids total des experts est de 5, comme celui des données observées. En particulier, $w_i^1 = 0.167$*,* $w_i^2 = 0.083$ *et* $w_i^3 = 0.056$*.*

En prenant en compte cette considération, si E experts totalement indépendants fournissent chacun $N(> n)$ pseudo-données pour lesquelles ils ont une certitude de 1, alors

- la somme des poids de l'expert e est de n/E et
- la somme des poids des experts est de n.

Ainsi, même dans le cas d'experts indépendants fournissant un grand nombre de données, le poids des pseudo-données n'excède pas le poids des observations. L'avis des experts *équivaut*, au maximum, à un jeu de données.

Poids des pseudo-données

Nous déterminons les poids des pseudo-données en prenant en compte la certitude des experts ainsi que les deux considérations (3.10) et (3.11). Nous posons donc

$$w_i^e = \frac{c_i^e}{1 + \sum_{j \neq i} r_{i,j}^2} \frac{n}{n_e \times E}. \tag{3.12}$$

Exemple 3. *Considérons le même contexte que dans l'exemple 2 avec des expert hautement dépendants* ($r_{i,j} = 1$)*. Si les poids* w_i^e *sont déterminés à partir de* (3.12)*, alors* $w_i^1 = 0.056$*,* $w_i^2 = 0.027$ *et* $w_i^3 = 0.018$*.*

3.3 *A priori* basé sur une pénalisation

Nous proposons dans ce qui suit, une approche alternative pour prendre en compte l'avis des experts dans le modèle. Nous nous intéressons en particulier ici directement à l'avis des experts concernant la fonction coefficient. Cette fonction coefficient du modèle de régression linéaire sur données fonctionnelles n'est pas simple à interpréter, même pour un statisticien. La nature fonctionnelle des données implique une certaine structure d'autocorrélation dans les données $x_i(\cdot)$. En particulier, la corrélation entre $x_i(t)$ et $x_i(t+h)$ est d'autant plus élevée que h est petit. Cette structure se retrouve dans une certaine mesure dans les quantités $x_i(\mathcal{I})$ si bien que, conditionnellement aux intervalles $\mathcal{I}$, le modèle Bliss est un modèle de régression linéaire multiple où les prédicteurs $(x_i(\mathcal{I}))$ sont corrélés (voir la section 2.2.2 du chapitre 2). Or, dans ce cas, les estimateurs admettent une grande variance si bien qu'il n'est pas évident de les interpréter (ce qui justifie les approches de régularisation comme *Ridge*, Hoerl and Kennard, 1970).
De plus, du fait de cette corrélation, il est possible qu'il y ait des artefacts d'estimation. En particulier, dans le cadre d'un modèle de régression linéaire multiple, si deux prédicteurs sont fortement corrélés, les coefficients de pente peuvent être mal estimés, l'un sur-estimé et l'autre sous-estimé. Dans le contexte du modèle Bliss, la corrélation des données peut induire la présence d'artefacts dans l'estimation de la fonction coefficient. Si bien que

la fonction coefficient est sur-estimée sur une période et sous-estimée sur une période adjacente.

Les experts sont tout de même capables de formuler des idées concernant une caractéristique de cette fonction. En particulier, conditionnellement à tout le reste, ils peuvent avoir un avis concernant :

1. Les périodes où la covariable fonctionnelle devrait avoir un impact (ou non) sur la variable réponse.

   Exemple 4. *Le niveau de précipitations sur une période donnée T influence la production de truffes. Au contraire, sur d'autres périodes, le niveau de précipitations n'impacte que de manière négligeable la production de truffes.*

2. Si l'impact est positif ou négatif sur la période T.

   Exemple 5. *Des précipitations plus élevées (resp. faibles) sur T que la précipitation moyenne sur T (observée sur les données) sont associées à une production de truffes élevée (resp. faible).*

L'avis des expert est donc sur le signe de la corrélation entre y et $x(t)$, pour t appartenant à une période donnée, conditionnellement aux les valeurs $x(t)$ sur les autres périodes. Pour modéliser ce type d'avis, nous définissons la fonction de signe $\beta^s(\cdot)$ par

$$\beta^s(t) = \mathbf{1}\left\{\beta(t) > 0\right\} - \mathbf{1}\left\{\beta(t) < 0\right\}, \tag{3.13}$$

qui correspond au signe de la fonction $\beta(\cdot)$.

Supposons éliciter de chaque expert e une fonction signe $\beta_e^s(\cdot)$ et une fonction de certitude $g_e(\cdot)$ telle que $g_e(t) \in [0,1]$. Pour prendre en compte l'avis des experts, nous proposons de modifier la distribution *a priori* de $(b, \mathcal{I})$ dans (3.2) donnée par

$$\pi_0(b, \mathcal{I}|\sigma^2) \propto |\Sigma(\mathcal{I})|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} b^T \Sigma(\mathcal{I})^{-1} b - a \sum_{k=1}^{K} \ell_k\right\}. \tag{3.14}$$

Les hyperparamètres b et $\mathcal{I}$ sont les paramètres relatifs à la fonction coefficient et donc à la fonction signe. La nouvelle distribution *a priori* que nous proposons est donnée par :

$$\pi(b, \mathcal{I}|\sigma^2; \tau) \propto \pi_0(b, \mathcal{I}|\sigma^2) \times \prod_{e=1}^{E} \exp\left\{-\tau \times \text{dist}^2(\beta^s, \beta_e^s; g_e)\right\}, \tag{3.15}$$

où $\pi_0(b, \mathcal{I}|\sigma^2)$ est donnée par (3.14), τ est un réel à fixer et dist est une distance entre la fonction signe β_e^s de l'expert e et la fonction signe β^s du modèle. De plus, cette distance doit prendre en compte la certitude g_e de l'expert. L'idée de cette modélisation est d'introduire un terme qui va pénaliser les fonctions coefficient $\beta(\cdot)$ qui ne sont pas *en accord* avec l'avis des experts. En choisissant (3.15) comme distribution *a priori*, la distribution *a posteriori* conditionnelle de $(b, \mathcal{I})$ est donnée par

$$\pi(b, \mathcal{I}|y, \sigma^2, \mu; \tau) \propto \exp\left\{-\left(\frac{1}{2\sigma^2}\text{SCR} + \tau \sum_{e=1}^{E} \text{dist}^2\left(\beta^s, \beta_e^s; g_e\right)\right)\right\} \times \pi_0(b, \mathcal{I}|\sigma^2). \tag{3.16}$$

Le terme $\tau\, \text{dist}^2\left(\beta^s, \bar{\beta}_E^s; \bar{g}_E\right)$ correspond alors à un terme de pénalisation. En effet, lorsqu'il augmente, il contribue à diminuer la valeur du terme exponentiel qui correspond ici à l'ajustement aux données. Ce terme de pénalisation fait intervenir :

- la distance dist qui sert à mesurer à quel point la fonction coefficient est *en accord* avec l'avis des experts et
- τ, un paramètre qui correspond à l'intensité de pénalisation.

Dans ce qui suit, nous expliquons quelle distance nous choisissons puis comment calibrer le paramètre τ.

3.3.1 Choix de la distance

Un choix possible est la distance L^2, pondérée par la fonction de certitude $g_e(\cdot)$, donnée par

$$\text{dist}^2(\beta^s, \beta_e^s; g_e) = \int_0^1 (\beta^s(t) - \beta_e^s(t))^2 g_e(t) \mathrm{d}t.$$

Dans ce cas, le terme de pénalisation dans (3.16) se réécrit comme la distance entre β^s et une fonction signe $\bar{\beta}_E^s$, étant la moyenne des β_e^s :

$$\pi(b, \mathcal{I}|\sigma^2; \tau) \propto \pi_0(b, \mathcal{I}|\sigma^2) \times \exp\left\{-\tau \times \text{dist}^2(\beta^s, \bar{\beta}_E^s; \bar{g}_E)\right\}$$

où $\bar{g}_E(t)$ est $\sum_{e=1}^E g_e(t)$ et

$$\bar{\beta}_E^s(t) = \sum_{e=1}^E \frac{g_e(t)}{\bar{g}_E(t)} \beta_e^s(t).$$

Ainsi, avec ce choix la distance entre la fonction signe β^s et les fonctions signe des experts se résument à une distance entre β^s et une fonction signe moyenne des experts. Ceci induit une simplification pour l'implémentation de cette méthode et diminue les temps de calculs.

Remarque. Dans (3.2), la distribution *a priori* $\pi_0(b, \mathcal{I}|\sigma^2)$ est définie afin d'être faiblement informative. Par conséquent, la valeur de la densité *a posteriori* de $(b, \mathcal{I})$ dans (3.16) est principalement déterminée par $\frac{1}{2\sigma^2}$SCR et le terme de pénalisation $\tau \,\text{dist}^2\left(\beta^s, \bar{\beta}_E^s; \bar{g}_E\right)$. Ainsi, si nous faisons augmenter la valeur de τ, la masse de la distribution *a posteriori* de $(b, \mathcal{I})$ est *tirée* d'une zone vraisemblable par rapport aux données, vers des valeurs de paramètres qui correspondent à des fonctions coefficient $\beta(\cdot)$ en accord avec l'avis des experts. Plus τ est grand, plus l'avis des experts est pris en compte dans la distribution *a posteriori*.

3.3.2 Calibration de τ

Concernant le paramètre de pénalisation τ, nous n'avons pas de connaissance *a priori* qui nous permette d'avoir une idée de sa valeur. Nous proposons deux manières de le fixer. La première se base sur une procédure de validation croisée bayésienne et la seconde consiste à mettre une distribution *a priori* sur τ.

Validation croisée bayésienne

Pour choisir τ, nous considérons dans cette section un ensemble de candidats : $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$. Nous écrivons le choix de la valeur de τ comme un choix de modèle parmi les modèles $\mathcal{M}_j$ pour $j = 1, \ldots, J$. Le modèle $\mathcal{M}_j$ est le modèle décrit en début de section 3.3 avec $\tau = \tau_j$. Pour en savoir plus sur les approches bayésiennes de sélection de modèle, on pourra consulter le chapitre 8 de Marin and Robert (2009) ou Chen et al. (2012).

L'approche que nous étudions consiste à reformuler le problème en termes de prise de décision, en introduisant une fonction d'utilité afin de "capturer l'utilité d'un modèle" (voir Gelfand et al., 1992).
Une manière de déterminer une utilité pour un modèle est de juger de sa qualité prédictive pour de nouvelles données $\tilde{z}$. Notons par $a(\tilde{z})$, une distribution prédictive de cette nouvelle observation. L'utilité introduite par Good (1952) est le score logarithmique, fonction de la prédictive a et de la future donnée $\tilde{z}$:

$$u(a, \tilde{z}) = \log a(\tilde{z}).$$

Ayant observé les données $\mathcal{D}$, pour un modèle $\mathcal{M}$ la meilleure distribution prédictive est celle maximisant l'espérance *a posteriori* de l'utilité

$$\begin{aligned}
\hat{a}(\tilde{z}) &= \arg\max_a \bar{u}(a; \mathcal{M}|\mathcal{D}) \\
&= \arg\max_a \int u(a, \tilde{z}) \pi(\tilde{z}|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\tilde{z} \\
&= \arg\max_a \int \log a(\tilde{z}) \, \pi(\tilde{z}|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\tilde{z} \\
&= \pi(\tilde{z}|\mathcal{D}, \mathcal{M}).
\end{aligned}$$

Pour comparer les modèles entre eux, nous comparons les meilleures distributions prédictives $\hat{a}(\tilde{z})$ de chacun des modèles en évaluant leurs utilités respectives. L'utilité d'un modèle $\mathcal{M}$ est donc définie comme l'utilité de sa meilleure prédictive :

$$\bar{u}(\mathcal{M}|\mathcal{D}) = \int \log \hat{a}(\tilde{z}) \, \pi(\tilde{z}|\mathcal{D}, \mathcal{M}) \, \mathrm{d}\tilde{z}, \tag{3.17}$$

et le modèle sélectionné sera celui ayant la plus grande utilité. Cependant, le calcul de (3.17) n'est pas forcément possible analytiquement. Une solution est d'utiliser une approximation Monte Carlo :

$$\bar{u}(\mathcal{M}|\mathcal{D}) \approx \frac{1}{N} \sum_{i=1}^{N} \log \pi(\tilde{z}_i|\mathcal{D}, \mathcal{M}),$$

où $\tilde{z}_1, \ldots, \tilde{z}_N$ est un échantillon suivant $\pi(\tilde{z}|\mathcal{D}, \mathcal{M})$. Pour avoir un tel échantillon, une astuce consiste à procéder comme pour une approche de validation croisée où pour un i fixé,

- z_i, une donnée de $\mathcal{D}$, joue le rôle d'une nouvelle observation $\tilde{z}$ à prédire et
- $\mathcal{D}_{-i}$, le jeu de données $\mathcal{D}$ privé de la i^e observation, joue le rôle des données d'apprentissage $\mathcal{D}$.

En prenant pour i des valeurs successives allant de 1 à n (nombre d'observations) on obtient un échantillon de $\tilde{z}$ (ce qui correspond à la validation croisée *Leave-One-Out*). Dans ce qui suit, nous détaillons la méthode que nous utilisons pour approcher l'utilité d'un modèle, détaillée dans Vehtari and Ojanen (2012). Dans notre contexte, les données $\mathcal{D}$ sont composées de n réplicats $z_i = \{y_i, x_i\}$ et l'approximation Monte Carlo est :

$$\bar{u}(\mathcal{M}_j|\mathcal{D}) \approx \bar{u}_{\text{LOO}}(\mathcal{M}_j|\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \log \pi(y_i|x_i, \mathcal{D}_{-i}, \mathcal{M}_j).$$

Pour calculer $\pi(y_i|x_i, \mathcal{D}_{-i}, \mathcal{M}_j)$, il faut intégrer sur le paramètre θ du modèle :

$$\pi(y_i|x_i, \mathcal{D}_{-i}, \mathcal{M}_j) = \int \pi(y_i|x_i, \mathcal{D}_{-i}, \theta, \mathcal{M}_j)\, \pi(\theta|\mathcal{D}_{-i}, \mathcal{M}_j)\, \mathrm{d}\theta. \tag{3.18}$$

Cette intégration n'étant pas possible dans certains contextes comme le notre, une première approche consisterait à faire une approximation MCMC de ce terme en se basant sur un échantillon de la loi *a posteriori* $\theta|\mathcal{D}_{-i}, \mathcal{M}_j$ pour chaque $i = 1, \ldots, n$. Cependant, nous préférerons une autre approche puisque le temps de calcul de cette approche est considérable. Pour approcher $\pi(y_i|x_i, \mathcal{D}_{-i}, \mathcal{M}_j)$, il est préférable d'utiliser une autre approche MCMC, un échantillonnage préférentiel où la loi d'importance est la loi *a posteriori* complète $\theta|\mathcal{D}, \mathcal{M}_j$:

$$\begin{aligned}\pi(y_i|x_i, \mathcal{D}_{-i}, \mathcal{M}_j) &= \mathbb{E}_{\theta|\mathcal{D}_{-i}, \mathcal{M}_j}\, \pi(y_i|x_i, \mathcal{D}_{-i}, \theta, \mathcal{M}_j) \\ &= \mathbb{E}_{\theta|\mathcal{D}, \mathcal{M}_j}\, \pi(y_i|x_i, \mathcal{D}_{-i}, \theta, \mathcal{M}_j) \frac{\pi(\theta|\mathcal{D}_{-i}, \mathcal{M}_j)}{\pi(\theta|\mathcal{D}, \mathcal{M}_j)}.\end{aligned}$$

Selon le modèle Bliss (3.1), y_i est indépendant de $\mathcal{D}_{-i}$ conditionnellement à x_i et à θ, donc $\pi(y_i|x_i, \mathcal{D}_{-i}, \theta, \mathcal{M}_j) = \pi(y_i|x_i, \theta, \mathcal{M}_j)$. En notant $\theta_1, \ldots, \theta_T$ un échantillon d'importance simulé suivant $\theta|\mathcal{D}, \mathcal{M}_j$, l'approximation est

$$\begin{aligned}\pi(y_i|x_i, \mathcal{D}_{-i}, \mathcal{M}_j) &\approx \sum_{t=1}^{T} \pi(y_i|x_i, \theta_t, \mathcal{M}_j) \frac{\pi(\theta_t|\mathcal{D}_{-i}, \mathcal{M}_j)}{\pi(\theta_t|\mathcal{D}, \mathcal{M}_j)} \\ &= \sum_{t=1}^{T} w_{-i,t} \times \pi(y_i|x_i, \theta_t, \mathcal{M}_j)\end{aligned}$$

où les poids d'importance $w_{-i,t}$ sont donnés par

$$w_{-i,t} = \frac{\pi(\theta_t|\mathcal{D}_{-i}, \mathcal{M}_j)}{\pi(\theta_t|\mathcal{D}, \mathcal{M}_j)}$$

et sont calculables numériquement en remarquant la simplification (par une utilisation du théorème de Bayes) :

$$w_{-i,t} = \frac{\pi(y_i|x_i, \mathcal{D}_{-i}, \mathcal{M}_j)}{\pi(y_i|x_i, \theta_t, \mathcal{M}_j)} \propto \frac{1}{\pi(y_i|x_i, \theta_t, \mathcal{M}_j)} \triangleq \tilde{w}_{-i,t}.$$

L'approximation de (3.18) est donnée par

$$\sum_{t=1}^{T} \frac{\tilde{w}_{-i,t}\, \pi(y_i|x_i, \theta_t, \mathcal{M}_j)}{\sum_{t=1}^{T} \tilde{w}_{-i,t}}$$

et comme $\tilde{w}_{-i,t}$ est l'inverse de $\pi(y_i|x_i, \theta_t, \mathcal{M}_j)$, elle se simplifie :

$$\frac{T}{\sum_{t=1}^{T} \pi(y_i|x_i, \theta_t, \mathcal{M}_j)^{-1}}.$$

L'utilité du modèle $\mathcal{M}_j$ est donc approchée par :

$$\bar{u}(\mathcal{M}_j|\mathcal{D}) \approx \bar{u}_{\text{IS-LOO}}(\mathcal{M}_j|\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{T}\sum_{t=1}^{T} \pi(y_i|x_i, \theta_t, \mathcal{M}_j)^{-1}\right).$$

Pour plus de détails concernant les approches bayésiennes de validation croisée, on pourra consulter Vehtari and Ojanen (2012).

Dans notre cas, nous approcherons l'utilité d'une valeur de $\tau_j \in \boldsymbol{\tau}$ par

$$\bar{u}_{\text{IS-LOO}}(\tau_j|\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{T}\sum_{t=1}^{T} \pi(y_i|x_i, \theta_t; \tau_j)^{-1}\right), \tag{3.19}$$

où $\theta_t = (\mu_t, b_t, \sigma_t^2, m_t, \ell_t)$ est obtenu en utilisant un algorithme d'échantillonnage (voir l'annexe 3.7.2) et θ_t suit la distribution *a posteriori* $\theta|\mathcal{D}; \tau_j$. Ici, les termes $\pi(y_i|x_i, \theta_t; \tau_j)$ sont simples à calculer. Nous sélectionnons donc la valeur de τ parmi $\boldsymbol{\tau}$ qui maximise (3.19).

A priori sur τ

Une manière de modéliser notre ignorance *a priori* sur la valeur de τ est de considérer cette quantité aléatoire et de lui attribuer une distribution *a priori* faiblement informative (voire non informative). Nous choisissons la loi *a priori* conjuguée

$$\tau|\lambda \sim \mathcal{E}(\lambda).$$

Par rapport à la valeur de λ, étant donnée que la distribution *a posteriori* de τ est proportionnelle à

$$\exp\left\{-\tau\left(\lambda + \text{dist}^2\left(\beta^s, \bar{\beta}_E^s; \bar{g}_E\right)\right)\right\} \tag{3.20}$$

un choix possible est de prendre $\lambda = 0$. Ce choix mène à une distribution dégénérée mais non informative dans le sens où l'expression de cette densité *a priori* n'apparaît dans la densité *a posteriori* qu'au travers d'une constante multiplicative.

3.4 Implémentation

Approche 1 : Pseudo-données Pour obtenir un échantillon de la distribution *a posteriori* $\pi(\theta|y, y^1, \ldots, y^E)$, nous utilisons une légère modification de l'échantillonneur de Gibbs décrit dans le chapitre 2. En effet, la modification du modèle est minime et les distributions conditionnelles sont quasiment les mêmes (voir Annexe 3.7.1).

Approche 2 : Pénalisation Pour cette approche, l'échantillonneur de la distribution *a posteriori* n'est pas le même que celui décrit dans le chapitre 2. En effet, l'ajout d'un terme de pénalisation, voir (3.15), implique que la loi *a priori* de b n'est plus conjuguée. Nous choisissons ici d'utiliser l'algorithme *Metropolis-Within-Gibbs*. A chaque itération, pour mettre à jour la valeur de b, nous proposons une marche aléatoire : on génère une valeur $b' = b+\varepsilon$ où $\varepsilon \sim \mathcal{N}_K(0, r\,\mathrm{Id}_K)$. La valeur proposée est acceptée ou refusée au regard d'une probabilité d'acceptation

$$\alpha = \min\left(1, \frac{\pi(b'|\mu, \sigma^2, m, \ell, \mathcal{D})}{\pi(b|\mu, \sigma^2, m, \ell, \mathcal{D})}\right),$$

de sorte que la distribution stationnaire de la chaîne de Markov soit la distribution *a posteriori* (voir Robert and Casella, 2013). Les autres étapes de mise à jour restent inchangées par rapport à l'algorithme précédent.
Une différence majeure est la nécessité de calibrer la valeur de r. Pour cela, nous exécutons plusieurs fois l'algorithme avec plusieurs valeurs de r. Nous gardons la valeur de r qui permet d'avoir un taux d'acceptation moyen entre 0.2 et 0.5. La valeur optimale pour un taux d'acceptation est 0.234 (Roberts and Rosenthal, 2001) dans le cas d'un modèle précis, mais au vu de la littérature, $[0.2, 0.5]$ semble un intervalle acceptable dans des modèles plus généraux. L'algorithme utilisé est décrit dans l'annexe 3.7.2.

3.5 Résultats numériques

3.5.1 Application sur des données simulées

Dans ce qui suit, les deux approches présentées en section 3.2 et 3.3 sont appliquées sur des données simulées, ce qui nous permet d'étudier leurs sensibilités par rapport à la certitude des experts. Pour simuler les données, nous utilisons le même schéma de simulation que celui présenté en section 2.3.1 du chapitre 2, avec les quantités suivantes :

- $\mu = 1$,
- $n = 100$,
- la fonction coefficient $\beta(\cdot)$ est de la forme *Step function* (voir Figure 2.2 du chapitre 2).
- σ^2 est tel que le ratio signal-bruit est de 5.

Les données obtenues sont considérées comme des données observées et sont notées $\mathcal{D}_0 = \{y_i, x_i(\cdot)\}_{i=1,\ldots,100}$.

Approche 1 : pseudo-données

Nous générons des pseudo-données que nous considérons issues de deux experts indépendants. Comme en section 3.2, nous considérons qu'en pratique les experts peuvent

prédire ou imaginer des données à partir de leurs connaissances. Nous faisons l'hypothèse que le modèle (3.4) est une approximation acceptable de leurs connaissances. Nous simulons 100 pseudo-données par expert suivant ce modèle avec une fonction coefficient spécifique à chaque expert. Les jeux de pseudo-données obtenus sont notés par $\mathcal{D}_1$ et $\mathcal{D}_2$. Nous considérons pour commencer que les experts attribuent une certitude 1 à chacune de leurs pseudo-données. Nous appliquons les méthodes sur ces jeux de données et les résultats graphiques sont présentés dans la figure 3.1.

Représentation des lois *a priori* et *a posteriori* Sur la figure 3.1 sont représentés les distributions de la fonction coefficient sachant les pseudo-données de l'expert 1 (graphique (a)), de l'expert 2 (graphique (b)) ou des experts 1 et 2 ensemble (graphique (c)). Même si nous représentons ici des distributions *a posteriori*, nous comprenons ces distributions comme une visualisation des *a priori* des experts puisqu'il s'agit de la distribution de $\beta(\cdot)$ sachant leurs pseudo-données. La distribution *a posteriori* de $\beta(\cdot)$ sachant les données observées $\mathcal{D}_0$ est donnée avec le graphique (d). Le graphique (e) montre la distribution *a posteriori* de $\beta(\cdot)$ sachant les données observées et sachant les pseudo-données de l'expert 1. Pour celle de l'expert 2, le résultat est donné avec le graphique (f). Le graphique (g) présente le résultat sachant les données observées et sachant toutes les pseudo-données. Ici, comme les certitudes des experts sont fixées à 1 et que les experts simulés sont considérés indépendants, la somme des poids des pseudo-données est égale au poids des données observées. Ainsi, pour les graphiques (e) (resp. (f)), la distribution *a posteriori* est également déterminée par l'expert 1 (resp. l'expert 2) et par les données observées. Pour le graphique (g), comme les pseudo-données des deux experts sont prises en compte, le poids de chacun des experts est la moitié du poids des données observées afin que le poids des pseudo-données ne dépasse pas celui des données observées.

Notons grâce à la propriété 3.2 que l'espérance *a posteriori* peut s'interpréter comme la moyenne entre l'espérance sachant les données observées et l'espérance sachant les pseudo-données. Les graphiques (e), (f) et (g) de la figure 3.1 illustrent cette propriété.

Sensibilité face aux certitudes Dans le paragraphe précédent, les certitudes étaient données égales à 1. Nous faisons maintenant varier ces certitudes de 1 vers 0. La figure 3.2 donne les distributions *a posteriori* pour différentes valeurs de certitudes, sachant $\mathcal{D}_1$ et $\mathcal{D}_2$. Le graphique (a) donne une illustration de la propriété 3.2, *i.e.* lorsque les certitudes c_i^e sont nulles, l'inférence ne dépend pas des pseudo-données. La distribution *a posteriori* obtenue est égale à la distribution *a posteriori* sachant seulement les données observées $\mathcal{D}_0$. Quand les certitudes augmentent (voir graphique (b) à (d)), l'espérance *a posteriori* se rapproche progressivement de l'espérance sachant les pseudo-données.

Approche 2 : pénalisation

Nous appliquons ici l'approche décrite en section 3.3 sur les données simulées $\mathcal{D}_0$, en ne gardant que 25 données. Nous ne gardons que peu de données pour être dans une situation où l'estimation est peu fiable. Nous pourrons alors évaluer si la prise en compte d'avis d'experts améliore ou non l'estimation. Nous nous donnons trois experts simulés

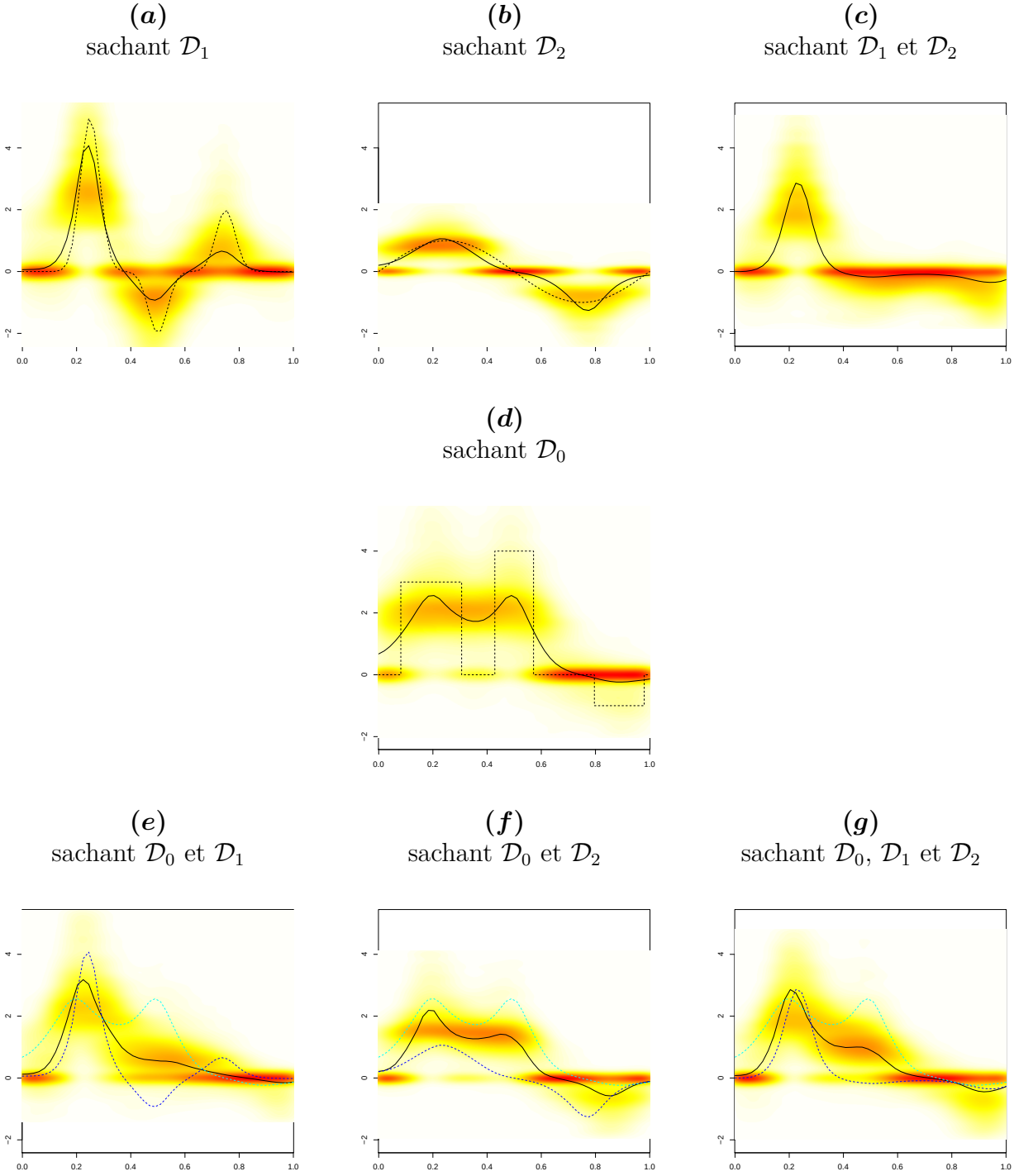


Figure 3.1: Distribution *a posteriori* de la fonction coefficient. *Les graphiques* (a), (b) *et* (c) *donnent la distribution a posteriori sachant les pseudo-données, le graphique* (d) *sachant les données observées et les graphiques* (e), (f) *et* (g) *sachant les données observées et sachant les pseudo-données. Pour chaque graphique, la courbe noire est l'espérance a posteriori. Pour le graphique* (a) *(resp.* (b) *et* (d)*), la courbe noire en pointillé est la fonction coefficient utilisée pour générer les données* $\mathcal{D}_1$ *(resp.* $\mathcal{D}_2$ *et* $\mathcal{D}_0$*). Pour les graphiques* (e), (f) *et* (g)*, la courbe cyan (resp. bleu) en pointillé est l'espérance sachant les données observées* $\mathcal{D}_0$ *(resp. les pseudo-données).*

avec des fonctions signes $\beta_e^s(\cdot)$ et des fonctions de certitudes $g_e(\cdot)$. La figure 3.3 résume l'avis des experts en présentant les fonctions $\bar{\beta}_E^s$ et $\bar{g}_E$.

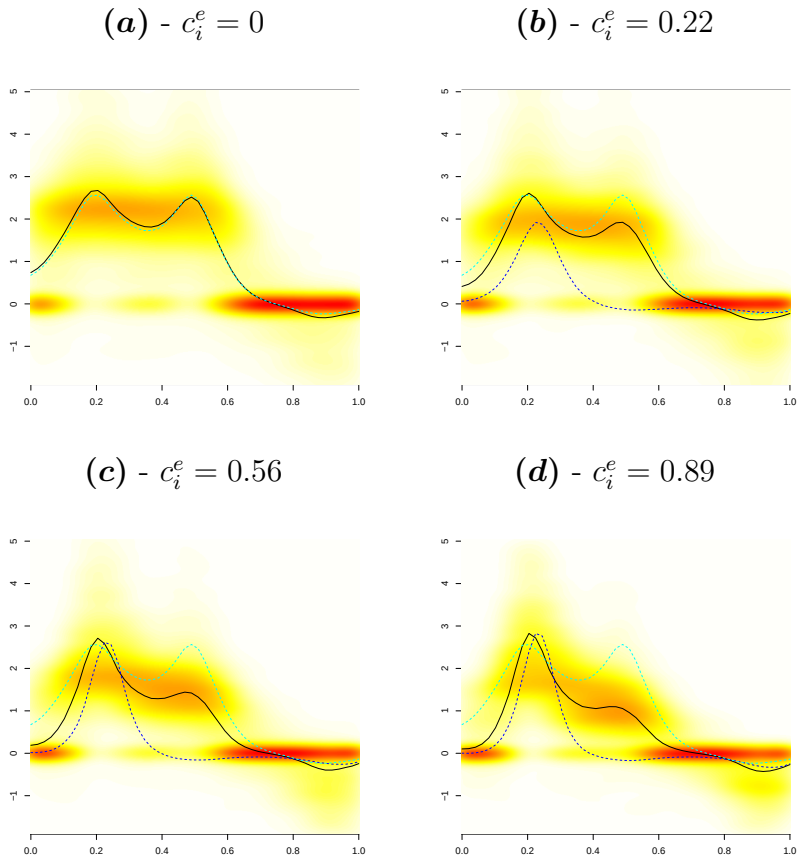


Figure 3.2: Exemples de distributions *a posteriori* pour différents niveaux de certitudes. *Chacun des graphiques represente la distribution a posteriori de la fonction coefficient sachant $\mathcal{D}_0$ (données observées) et $\mathcal{D}_1$, $\mathcal{D}_2$ (pseudo-données). La courbe noire en trait plein est l'espérance a posteriori et la courbe cyan (resp. bleu) en pointillé est l'espérance sachant les données observées (resp. les pseudo-données).*

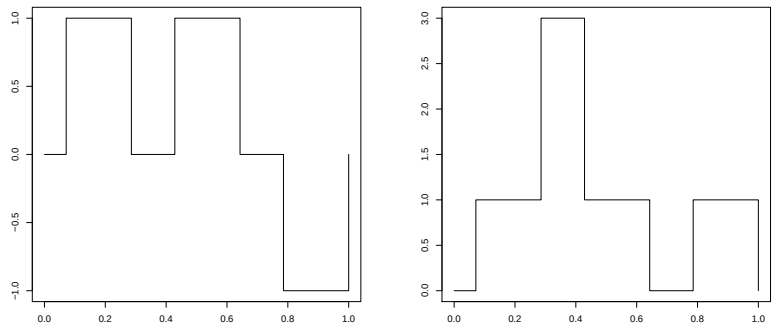


Figure 3.3: Représentation de la connaissance des experts (simulés) en utilisant la procédure d'élicitation décrite en section 3.3. *Le premier (resp. second) graphique donne la fonction $\bar{\beta}_E^s$ (resp. $\bar{g}_E$), la fonction signe moyenne des experts (resp. la certitude totale des experts).*

Validation croisée bayésienne Pour calibrer l'hyperparamètre τ, voir (3.15), nous utilisons la procédure de la validation croisée bayésienne décrite en section 3.3.2. Nous sommes amenés à choisir une valeur de τ parmi un ensemble $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$ qu'il faut

se donner. Cependant, il n'est pas évident de se donner un tel ensemble. Nous expliquons dans ce qui suit comment nous fixons $\boldsymbol{\tau}$ et comment nous choisissons la valeur de τ.

1. Fixer l'ensemble $\boldsymbol{\tau}$ Pour commencer, constatons que seules les valeurs positives ont un sens par rapport à la notion de pénalité. De plus, si $\tau = 0$, le terme de pénalité n'apparaît pas, autrement dit cela revient à ne pas prendre en compte l'avis des experts. Nous choisissons de fixer à 0 la plus petite valeur de $\boldsymbol{\tau}$. Pour choisir la plus grande valeur de $\boldsymbol{\tau}$, rappelons la distribution *a posteriori* des paramètres de la fonction coefficient :

$$\pi(b, \mathcal{I}|y, \mu, \sigma^2; \tau) \propto \exp\left\{-\frac{1}{2\sigma^2}\text{SCR} - \tau \times \text{dist}^2(\beta^s, \bar{\beta}^s_E; \bar{g}_E)\right\} \pi_0(b, \mathcal{I}|\sigma^2).$$

Si la valeur de τ est trop petite, alors le terme de pénalisation sera négligeable devant SCR et l'avis des experts n'aura aucun poids dans l'analyse. A l'opposé, si la valeur de τ est trop grande, le terme de pénalisation sera trop important face à SCR et les données ne compteront pas. L'objectif est donc de proposer un ensemble $\boldsymbol{\tau}$ dont la plus grande valeur τ_N soit telle que $\frac{1}{2\sigma^2}\text{SCR}$ et $\tau_N \,\text{dist}^2(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$ soient comparables.
Pour déterminer τ_N, nous avons besoin de connaître l'amplitude des quantités aléatoires $\frac{1}{2\sigma^2}\text{SCR}$ et $\text{dist}^2(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$. Nous les approchons en utilisant un échantillon *a posteriori*, avec τ fixé à 0. Notons par $\hat{\mathbb{E}}\left(\text{SCR}(\sigma^2)\right)$ la moyenne des valeurs du premier terme et par $\hat{\mathbb{E}}\left(\text{dist}\right)$ celle du second. Nous fixons alors τ_N comme

$$\tau_N = \frac{\hat{\mathbb{E}}\left(\text{SCR}(\sigma^2)\right)}{\hat{\mathbb{E}}\left(\text{dist}\right)}. \tag{3.21}$$

Les distributions de $\frac{1}{2\sigma^2}\text{SCR}$ et $\text{dist}^2(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$ sont données dans la figure 3.4 et τ_N est fixé à 16.444. L'ensemble $\boldsymbol{\tau}$ est donc fixé comme une grille régulière de 50 valeurs allant de 0 à 16.444.

2. Calcul des utilités Pour calculer les valeurs $\bar{u}_{\text{IS-LOO}}(\tau)$ pour toutes les valeurs de $\tau \in \boldsymbol{\tau}$, nous utilisons l'approximation (3.19). Le nombre T correspond à la taille de l'échantillon d'importance utilisé pour calculer une approximation Monte Carlo des poids d'importance $w_{-i,t}$. Nous choisissons $T = 10000$ et nous représentons avec la figure 3.5 les approximations de $\bar{u}_{\text{IS-LOO}}(\tau)$ pour $\tau \in \boldsymbol{\tau}$.

3. Sélection de la valeur de τ Nous choisissons la valeur de τ qui maximise l'utilité approchée (3.19) décrite en section 3.3.2. Nous choisissons ici $\tau = 16.115$. Nous donnons avec le second graphique de la figure 3.5, les valeurs de $\bar{u}_{\text{IS-LOO}}(\tau)$ pour τ allant de 0 à $2 \times \tau_N$. On constate que pour des valeurs de τ plus grande que τ_N, l'utilité $\bar{u}_{\text{IS-LOO}}(\tau)$ se stabilise, ce qui confirme, empiriquement, le choix de $\tau = 16.115$. On obtient ce type de résultat pour l'utilité lorsque l'information apportée par les experts est en accord avec le schéma de génération des données observées. Lorsque ça n'est pas le cas, l'utilité du modèle diminue lorsque τ augmente. Pour illustrer cela, on donne avec la figure 3.6, les valeurs de $\bar{u}_{\text{IS-LOO}}(\tau)$ lorsque la fonction coefficient utilisée pour générer les données observées est de la forme *Smooth* (voir Figure 2.2 du chapitre 2).

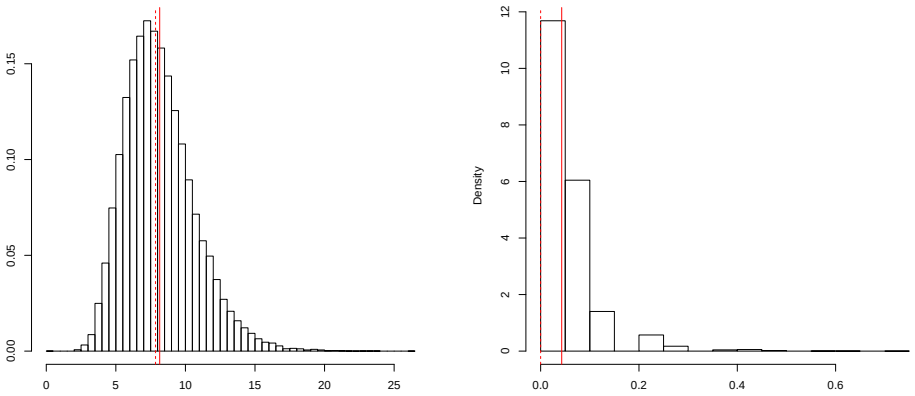


Figure 3.4: Distributions *a posteriori* des termes $\frac{1}{2\sigma^2}$SCR et dist$(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$ pour des données simulées lorsque τ est fixé par validation croisée. *Le premier (resp. second) graphique représente la distribution a posteriori de $\frac{1}{2\sigma^2}SCR$ (resp. de* dist$(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$*). La droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution.*

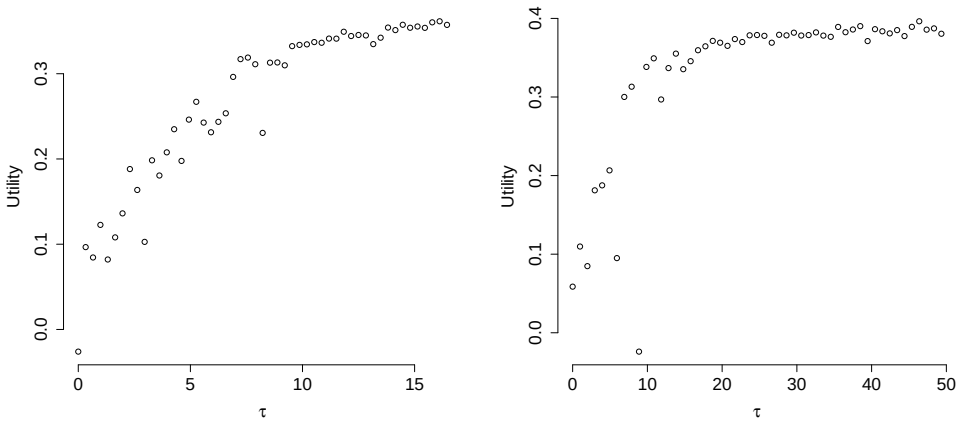


Figure 3.5: Résultats numériques pour différentes valeurs de τ. *Le premier graphique donne les approximations numériques des utilités $u_{IS\text{-}LOO}(\tau)$ pour chaque valeur de τ dans $\boldsymbol{\tau}$. Le second graphique donne les approximations de l'utilité pour les valeurs de τ allant de 0 à $2 \times \tau_N$.*

Rappelons que dans cette section nous considérons un jeu de données avec $n = 25$ observations. De plus, l'avis des experts représenté dans la figure 3.3 correspond à la vraie fonction coefficient qui est de la forme *Step function* (voir la figure 2.2 du chapitre 2). La figure 3.7 montre la distribution *a posteriori* de la fonction coefficient quand $\tau = 0$ et quand τ est calibré par validation croisée bayésienne. Nous constatons que l'estimation de la fonction coefficient est assez mauvaise lorsque l'avis des experts n'est pas pris en compte (le premier graphique de la figure 3.7). Ce résultat n'est pas étonnant puisque le nombre d'observations est ici assez faible. Avec le second graphique, nous pouvons constater l'apport de la prise en compte de l'avis des experts. La distribution *a posteriori* et l'estimation sont alors plus précises.

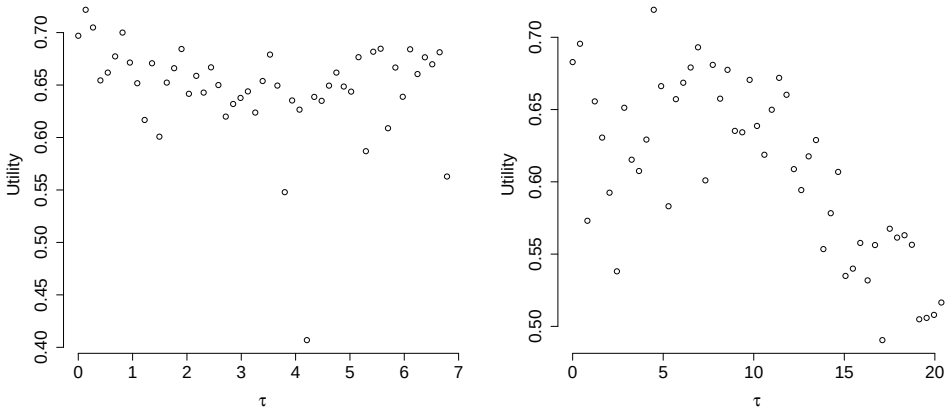


Figure 3.6: Résultats numériques pour différentes valeurs de τ lorsque la vraie fonction coefficient est de la forme *Smooth*. *Le premier graphique donne les approximations numériques des utilités $u_{IS\text{-}LOO}(\tau)$ pour chaque valeur de τ dans $\boldsymbol{\tau}$. Le second graphique donne les approximations de l'utilité pour les valeurs de τ allant de 0 à $2 \times \tau_N$.*

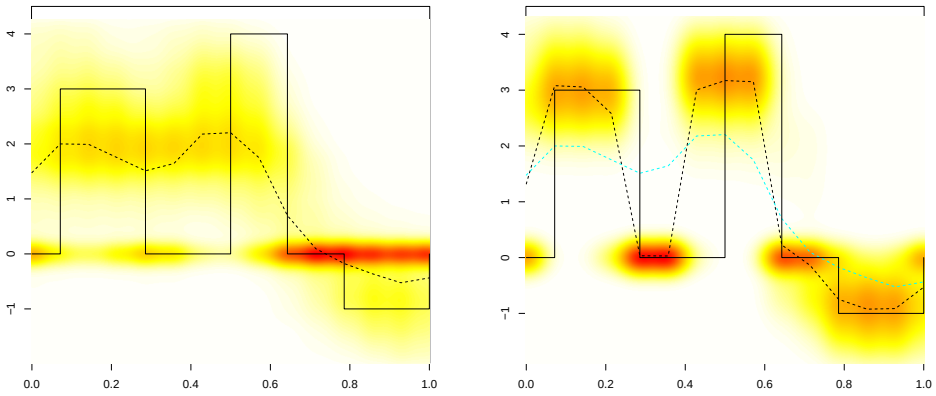


Figure 3.7: Distribution *a posteriori* de la fonction coefficient pour différents valeurs de τ, pour τ fixé par validation croisée. *Le premier (resp. second) graphique représente la distribution a posteriori de la fonction coefficient quand $\tau = 0$ (resp. $\tau = 16.115$), ce qui correspond au cas où aucune connaissance d'expert n'est prise en compte. La courbe noire en trait plein est la fonction coefficient "Step function" utilisée pour générer les données. La courbe noire en pointillé est l'espérance a posteriori. La courbe cyan en pointillé pour le second graphique est l'espérance a posteriori quand $\tau = 0$.*

***A priori* sur τ** Nous appliquons maintenant l'approche décrite en section 3.3.2 dans laquelle τ est considéré comme aléatoire. Nous choisissons de prendre $\lambda = 0$ si bien que $\tau \sim \mathcal{E}(0)$ est une loi *a priori* non-informative.

La figure 3.8 montre les distributions *a posteriori* de la fonction coefficient et de τ. Le premier graphique permet de constater qu'il n'y a pas de différence majeure entre l'analyse avec ou sans avis d'experts. Le second graphique montre que la distribution *a posteriori* de τ est globalement concentrée sur l'intervalle $[0, 10]$. Ces valeurs sont assez faibles par rapport à la valeur choisie par validation croisée (16.115). Nous donnons aussi dans la figure 3.9 les distributions *a posteriori* des termes $\frac{1}{2\sigma^2}\text{SCR}$ et $\tau \times \text{dist}(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$.

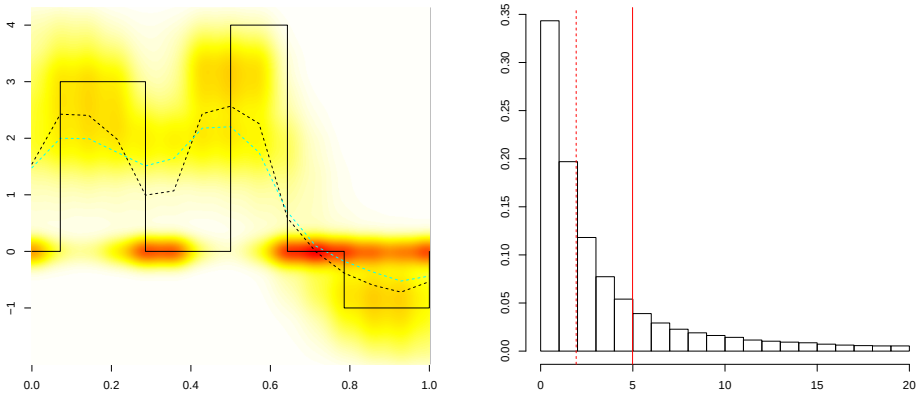


Figure 3.8: Distributions *a posteriori* lorsque $\tau \sim \mathcal{E}(0)$. *Le premier (resp. second) graphique représente la distribution a posteriori de la fonction coefficient (resp. de τ). Pour le premier graphique, la courbe noire en pointillé est l'espérance et celle en bleu est l'espérance si on fixe $\tau = 0$. Pour le second graphique, la droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution représentée.*

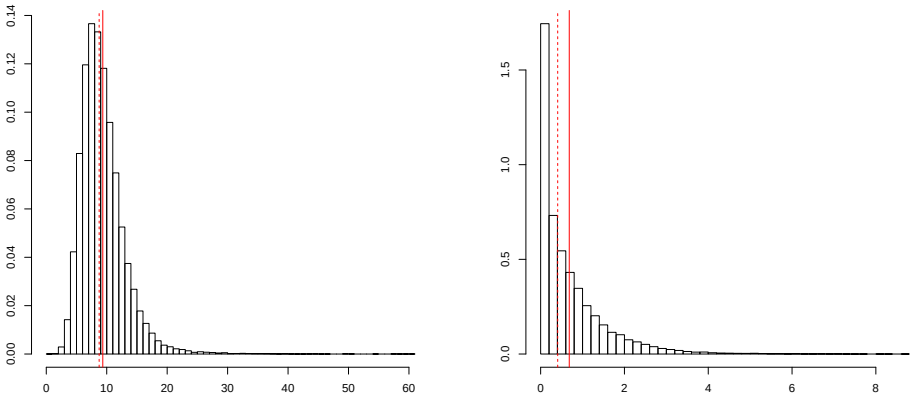


Figure 3.9: Distributions *a posteriori* des termes $\frac{1}{2\sigma^2}$SCR et $\tau \times \text{dist}(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$ pour des données simulées lorsque $\tau \sim \mathcal{E}(0)$. *Le premier (resp. second) graphique représente la distribution a posteriori de $\frac{1}{2\sigma^2}SCR$ (resp. de $\tau \times \text{dist}(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$). La droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution représentée.*

Cela nous permet de constater que le terme relatif à l'avis des experts est trop faible par rapport à l'ajustement aux données observées : $\frac{1}{2\sigma^2}$SCR. L'avis des experts n'est donc pas pris suffisamment en compte du fait des faibles valeurs de τ. Ceci explique pourquoi l'estimation de la fonction coefficient n'est pas très différente lorsque nous prenons en compte l'avis des experts.

3.5.2 Application sur les données de truffes noires du Périgord

Dans le chapitre 2, l'approche Bliss est appliquée sur des données agronomiques recueillant des productions de truffes noires du Périgord et des précipitations durant le

cycle de vie des truffes. L'objectif de cette application est de comprendre l'influence des précipitations sur la production de truffes afin d'aider les trufficulteurs dans la gestion des truffières mais aussi d'appréhender les effets du changement climatique. Les mesures de productions de truffes sont difficiles à obtenir et leur collecte constitue un travail non négligeable, ce qui explique pourquoi le nombre de données disponibles est peu important. Dans ce qui suit, nous étudions les données d'une truffière de Pernes-Les -Fontaines (Vaucluse, France) fournies par J. Gravier. Nous disposons de 25 années de productions allant de la saison 1914-1915 à la saison 1948-1949. Les mesures de précipitations commencent en Janvier d'une année n et finissent en Mars de l'année $n+1$. Les observations de précipitations correspondent aux cumuls des pluies pendant un mois, si bien que la cinétique de précipitations est résumée par 15 mesures mensuelles.
Le but de cette section est de pallier le manque de données en prenant en compte les connaissances des experts afin de renforcer l'inférence statistique. Comme décrit dans le chapitre 2, les études concernant la croissance et les mécanismes de reproduction de la truffe (Le Tacon et al., 2014, 2016) permettent d'émettre l'hypothèse que la production de truffes serait principalement sensible aux conditions climatiques sur certaines périodes. Les périodes critiques seraient l'été et la fin de l'hiver de la seconde année.

Pour cette étude, nous avons collaboré avec 3 types d'experts en trufficulture : des chercheurs, des trufficulteurs et le président de la Fédération Française des Trufficulteurs (FFT). Nous avons obtenu l'avis de F. Le Tacon, C. Murat, P. Montpied (chercheurs à l'INRA, Nancy, France), Joël Gravier, Pierre Cunty (trufficulteurs) et Michel Tournayre (président de la FFT). Nous leur avons introduit les idées principales des approches bayésiennes, avec en particulier l'intérêt d'une distribution *a priori* informée par leurs connaissances puis nous avons appliqué les méthodes proposées en section 3.2 et 3.3 pour intégrer leurs avis dans le modèle Bliss.

Approche 1 : pseudo-données

Élicitation Pour obtenir des pseudo-données, nous avons proposé aux experts un formulaire dont une première partie est présentée dans la figure 3.10. Le formulaire en comportait plusieurs copies afin que les experts puissent être libres de renseigner tous les scénarios auxquels ils pensaient. Globalement, les experts ont pu renseigner 2 ou 3 scénarios différents, excepté un trufficulteur qui en a renseigné 6. Il n'était pas évident pour eux de penser à beaucoup de situations de précipitations et de productions. Pour ces experts, les pseudo-données ainsi renseignées correspondaient à des scénarios typiques. Or ces scénarios typiques ne permettent pas d'appréhender l'ensemble de leurs connaissances. Dans le but d'extraire le plus de connaissances possibles, nous avons proposé aux experts un cadre plus simple, voir la figure 3.11. Nous leur avons fourni des scénarios de précipitations et il leur était demandé quelle devrait être vraisemblablement la production de truffes. Les scénarios de précipitations proposées ont été piochés dans un jeu de données d'une étude précédente similaire, inconnue des experts. Cette partie du questionnaire nous a permis de récolter 20 pseudo-données supplémentaires par expert.

Poids des pseudo-données Dans notre cas, trois experts (F. Le Tacon, C. Murat et P. Montpied) travaillent dans la même équipe à l'INRA de Nancy. Pour prendre en

Renseigner un scénario de précipitations, une production vraisemblable de truffes et votre certitude en votre avis.

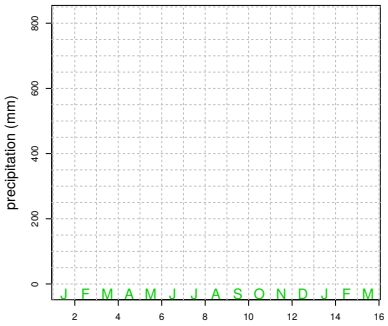
Courbe de précipitations	**Précipitations en mm (par mois)**	**Production plausible de truffes (en kilo par hectare)**
	Janvier : Février : Mars : Avril : Mai : Juin : Juillet : Août : Septembre : Octobre : Novembre : Décembre : Janvier : Février : Mars :	**Indice de certitude**

Figure 3.10: Une partie du questionnaire donné aux experts. (1) *Les experts doivent donner un scénario de précipitations, une production de truffes vraisemblable, et leur certitude en leur avis.*

compte cette proximité et le fait que leur avis devraient être similaires, nous fixons une dépendance forte (0.8) entre les pseudo-données de ces trois experts. La table 3.2 donne des exemples de poids w_i^e que nous obtenons à partir des certitudes des experts et en prenant en compte les corrections détaillées en section 3.2.3.

Par exemple, F. Le Tacon (resp. M. Tournayre) a donné une certitude de 0.8 concernant sa première (resp. deuxième) pseudo-donnée. Rappelons que $n = 25$, $E = 6$, que F. Le Tacon et M. Tournayre ont respectivement fourni 23 et 22 pseudo-données et que le coefficient de dépendance entre F. Le Tacon, C. Murat et P. Montpied est fixé à 0.8. En utilisant (3.12), les poids w_1^1 (pour F. Le Tacon) et w_2^6 (pour M. Tournayre) sont :

$$w_1^1 = \frac{c_1^1}{1 + 2 \times 0.8^2} \times \frac{25}{23 \times 6} = 0.06356471 \qquad w_2^6 = c_2^6 \times \frac{25}{22 \times 6} = 0.1515152.$$

Cet exemple permet de constater l'impact de la prise en compte de la dépendance entre certains experts (voir (3.10)) puisque F. Le Tacon est considéré fortement dépendant avec deux experts alors que M. Tournayre est indépendant des autres experts.
Comparons maintenant les poids obtenus pour les premières pseudo-données de J. Gravier et M. Tournayre. Ces deux experts sont considérés indépendants et ont respectivement fourni 26 et 22 pseudo-données. Ces deux experts ont donné une certitude de 1 pour leurs premières pseudo-données, d'où

$$w_1^3 = c_3^1 \times \frac{25}{26 \times 6} = 0.1602564 \qquad w_1^6 = c_1^6 \times \frac{25}{22 \times 6} = 0.1893939.$$

Cet exemple permet de constater l'impact de la prise en compte du nombre de pseudo-données fournies par chaque expert (voir (3.11)).

Pour le scénario de précipitations donné, donner une production vraisemblable de truffes et votre certitude en votre avis.

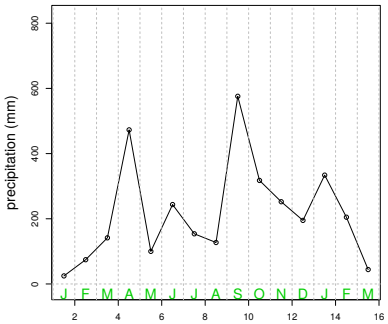
Courbe de précipitations	**Précipitations en mm (par mois)**	**Production plausible de truffes (en kilo par hectare)**
	Janvier : 24.7 Février : 74.4 Mars : 141.5 Avril : 472.8 Mai : 100.0 Juin : 243.6 Juillet : 153.9 Août : 127.2 Septembre : 575.8 Octobre : 317.6 Novembre : 252.3 Décembre : 195.0 Janvier : 333.9 Février : 204.8 Mars : 44.2	**Indice de certitude**

Figure 3.11: Une partie du questionnaire donné aux experts. (2) *Pour un scénario de précipitations donné, les experts doivent donner une production de truffes vraisemblable et leur certitude en leur avis.*

L'expression (3.8) de la distribution *a posteriori* permet de constater que la somme des poids des pseudo-données $\sum_{e=1}^{E} n_e w^e = \sum_{e=1}^{E} \sum_{i=1}^{n_e} w_i^e$ joue le même rôle que la taille d'échantillon n. En ce sens, nous interprétons la somme des poids d'un expert comme un équivalent échantillon, *i.e.* l'avis de l'expert e compte dans l'inférence comme $\sum_{i=1}^{n_e} w_i^e = n_e w^e$ données. Par exemple, l'avis de J. Gravier compte comme 3.734 données, voir la table 3.2.

Table 3.2: ***Un exemple de 5 certitudes et leurs poids associés.*** *Les poids sont calculés par* (3.12) *à partir des certitudes des experts. Une dépendance forte* (0.8) *entre F. Le Tacon, C. Murat et P. Montpied a été considérée. Les sommes des poids de chaque expert e sont données par* $n_e w^e$.

F. Le Tacon		C. Murat		J. Gravier		P. Montpied		P. Cunty		M.Tournayre	
$n_1 w^1 = 0.882$		$n_2 w^2 = 0.628$		$n_3 w^3 = 3.734$		$n_4 w^4 = 0.091$		$n_5 w^5 = 2.159$		$n_6 w^6 = 3.447$	
c_i^1	w_i^1	c_i^2	w_i^2	c_i^3	w_i^3	c_i^4	w_i^4	c_i^5	w_i^5	c_i^6	w_i^6
0.8	0.064	0.5	0.04	1	0.16	0.05	0.005	0	0	1	0.189
0.7	0.056	0.5	0.04	1	0.16	0.05	0.005	0.5	0.095	0.8	0.152
0.5	0.04	1	0.079	0.8	0.128	0.05	0.005	0.3	0.057	0.8	0.152
0.3	0.024	0.5	0.04	1	0.16	0.05	0.005	0.5	0.095	0.8	0.152
0.6	0.048	0.5	0.04	1	0.16	0.05	0.005	0.3	0.057	0.7	0.133

Résultats La figure 3.12 présente les résultats que nous obtenons sur les données de Pernes-Les-Fontaines et en prenant en compte les pseudo-données des 6 experts.

Le graphique (*a*) montre la distribution *a posteriori* de la fonction coefficient sachant les données observées. Les données mettent en lumière une période positive pendant l'été et dans une moindre mesure une période négative pendant l'automne.
Le graphique (*b*) montre la distribution *a posteriori* de la fonction coefficient sachant les pseudo-données, qui correspond à l'avis *a priori* des experts. On trouve une période positive pour les mois d'été et une période négative en hiver, ce qui correspond aux avis que les experts pouvaient formuler oralement. En un certain sens, cela indique que cette approche permet de collecter fidèlement l'avis des experts sous forme de pseudo-données.
Le graphique (*c*) donne la distribution *a posteriori* de la fonction coefficient sachant les données observées et les pseudo-données. En analysant les différences entre la courbe noire en trait plein et la courbe cyan en pointillé, on peut appréhender l'apport de l'avis des experts. Dans un premier temps, on constate que la période détectée en été n'est guère modifiée. En un certain sens, les données observées étaient suffisantes pour bien détecter l'effet positif des précipitations en été sur la production de truffes. Par contre, dans le cas de la second période (effet négatif en hiver), on constate que l'avis des experts a induit un léger changement dans l'estimation de la fonction coefficient. En effet, l'estimation est plus élevé (en valeur absolue) et l'effet s'étale sur une période plus grande, jusqu'à la fin de l'hiver.

Approche 2 : pénalisation

Élicitation On illustre maintenant sur les données de truffes la deuxième approche décrite en section 3.3. Nous demandons à chaque expert de donner son avis sur les périodes du cycle de vie de la truffe pendant lesquelles les précipitions ont un impact important sur la production de truffes. Ils doivent aussi renseigner si selon eux l'effet est positif ou négatif sur ces périodes. Autrement dit, cela revient à réfléchir à la question suivante : "Si pour la période renseignée on observe un niveau de précipitations plus élevé (ou moins élevé) que les précipitations moyennes sur cette même période les autres années, cela induit-il (toutes choses égales par ailleurs) une hausse ou une chute de la production de truffe ?" De plus, pour chaque paire période/effet renseignée, leur certitude (entre 0 et 1) leur est demandée.

La figure 3.13 donne la partie du formulaire que nous avons donné aux experts afin qu'ils renseignent leur avis sur les caractéristiques de la fonction coefficient. À partir des réponses à ce formulaire, nous avons encodé leurs fonctions signe dont nous présentons un résumé avec la figure 3.14. Cette figure donne la fonction signe moyenne des experts $\bar{\beta}_E^s$ et la certitude globale des experts $\bar{g}_E$. Le premier graphique de cette figure permet de représenter l'avis des experts : ils considèrent globalement que les précipitations en été devraient avoir un effet positif sur la production de truffes alors qu'en hiver elles devraient avoir un effet négatif. On constate donc que la fonction signe moyenne $\bar{\beta}_E^s$ est en accord avec ce que nous avons obtenu dans la section précédente, voir le graphique (*b*) de la figure 3.12.

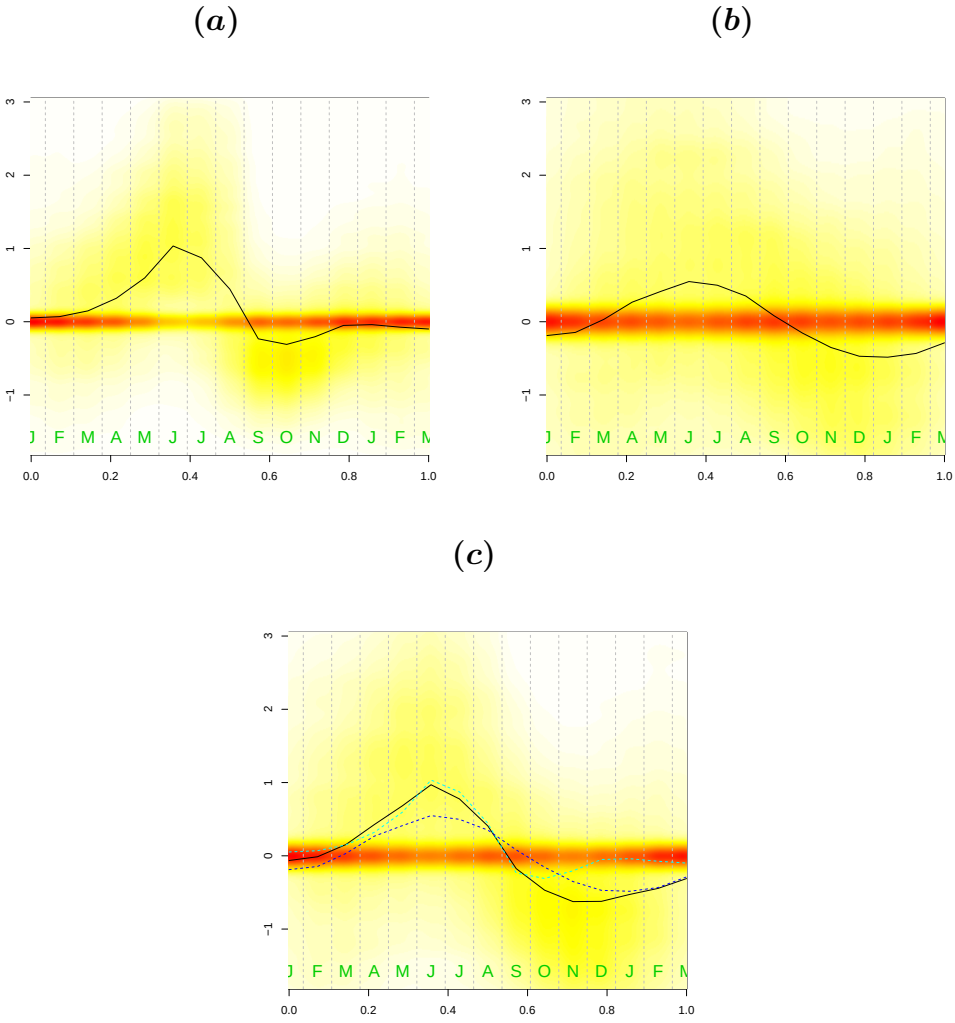


Figure 3.12: Résultats obtenus pour les données de truffes en prenant en compte les pseudo-données des experts. *Le graphique* (a) *(resp.* (b)*) représente la distribution a posteriori de la fonction coefficient sachant les données observées (resp. les pseudo-données). Le graphique* (c) *montre la distribution a posteriori sachant les données observées et sachant les pseudo-données. Pour chaque graphique, la courbe noire en trait plein correspond à l'espérance associée à la distribution représentée. La courbe en pointillé cyan (resp. bleu) est l'espérance de la fonction coefficient sachant les données observées (resp. les pseudo-données).*

Résultat en utilisant la validation croisée bayésienne Nous considérons dans un premier temps, l'approche où τ est calibré en utilisant une procédure de validation croisée. En suivant la même procédure que celle utilisée en section 3.5.1 pour l'étude sur des données simulées, nous fixons une grille $\boldsymbol{\tau}$ allant de 0 à 5.307. Le premier graphique de la figure 3.15 représente les valeurs de l'utilité approximée (3.19) pour toutes les valeurs de $\tau \in \boldsymbol{\tau}$. Nous sélectionnons $\tau = 4.988$. Le second graphique de la figure 3.15 donne la distribution *a posteriori* de la fonction coefficient. Notons que la prise en compte de l'avis des experts avec cette valeur de τ n'induit qu'une différence mineure dans l'estimation de la fonction coefficient. Pour comprendre pourquoi il y a si peu de différence, nous donnons dans la figure 3.16 les distributions *a posteriori* de $\frac{1}{2\sigma^2}\text{SCR}$ et de $\tau \times \text{dist}^2(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$. On constate que le terme de pénalisation et relativement faible par rapport à la somme des carrés des résidus.

Donner les périodes pour lesquelles les précipitations ont un impact sur la production de truffes, le type d'impact et votre certitude.

Période 1 :	Impact :	certitude :
Période 2 :	Impact :	certitude :
Période 3 :	Impact :	certitude :
Période 4 :	Impact :	certitude :
Période 5 :	Impact :	certitude :
Période 6 :	Impact :	certitude :
Période 7 :	Impact :	certitude :

Figure 3.13: Une partie du questionnaire donné aux experts. (3)

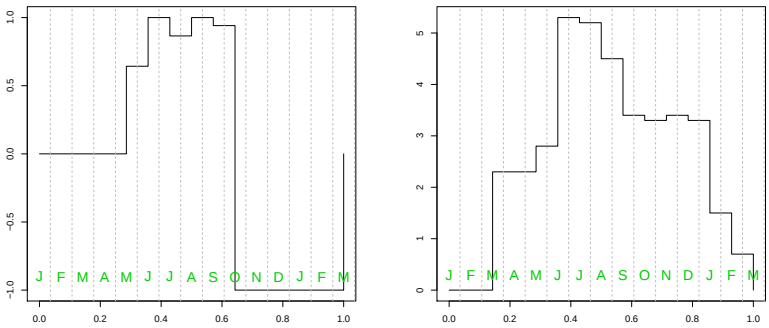


Figure 3.14: Résumé des avis des experts concernant le support et le signe de la fonction coefficient. *Le premier graphique montre la fonction signe moyenne des experts* $\bar{\beta}_E^s$ *et le second donne leur certitude globale* $\bar{g}_E$.

Résultat en utilisant une loi *a priori* sur τ Nous considérons maintenant le cas où τ suit *a priori* une loi exponentielle. Les résultats obtenus sont présentés avec la figure 3.17 qui donne la distribution *a posteriori* de $\beta(\cdot)$. Pour savoir à quel point les avis d'experts ont compté, nous calculons à partir d'un échantillon MCMC les distributions *a posteriori* de $\frac{1}{2\sigma^2}\text{SCR}$ et de $\tau \times \text{dist}^2(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$, qui sont données dans la figure 3.17. Notons avec les graphiques (c) et (d) que le terme de pénalisation (globalement entre 0 et 2) n'est pas du même ordre de grandeur que le terme $\frac{1}{2\sigma^2}\text{SCR}$ (globalement entre 10 et 20). Ce dernier point explique pourquoi on ne constate pas de différence marquante entre l'espérance *a posteriori* avec $\tau = 0$ et celle où $\tau \sim \mathcal{E}(0)$.

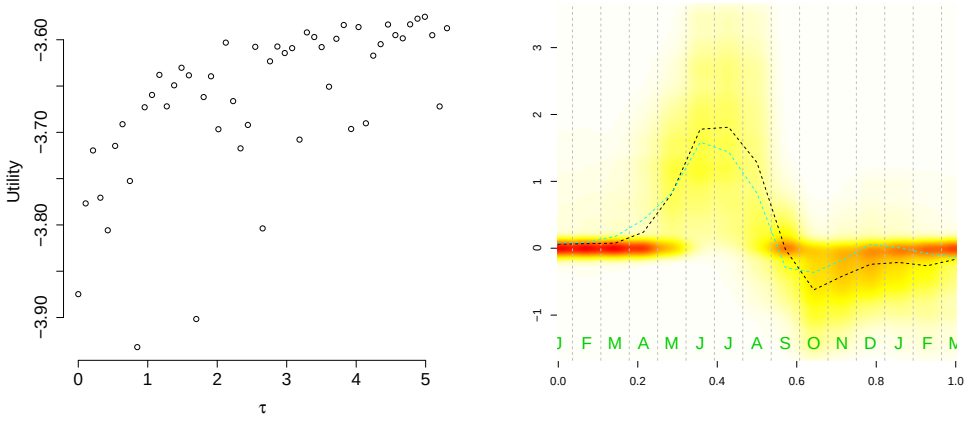


Figure 3.15: Résultats numériques pour les données de truffes quand τ est calibré par validation croisée bayésienne. *Le premier graphique donne les valeurs du critère de validation croisée pour chaque valeur de $\boldsymbol{\tau}$, voir* (3.19)*. Le second graphique montre la distribution a posteriori de la fonction coefficient.*

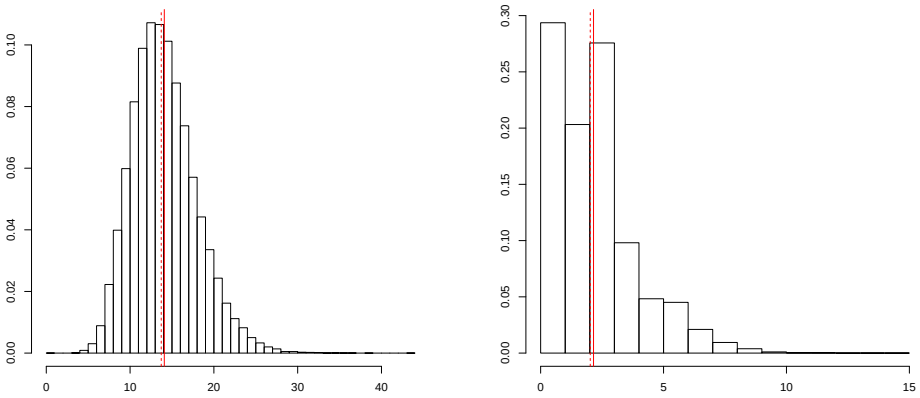


Figure 3.16: Distribution *a posteriori* de $\frac{1}{2\sigma^2}$SCR et de $\tau \times \mathbf{dist}^2(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$ pour les données de Pernes-Les-Fontaines (τ fixé par validation croisée bayésienne). *Le premier (resp. second) graphique est un histogramme des valeurs de $\frac{1}{2\sigma^2}SCR$ (resp. de $\tau \times \text{dist}^2(\beta^s, \bar{\beta}_E^s; \bar{g}_E)$) calculé à partir de l'échantillon MCMC. Pour ces deux graphiques, la droite rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution représentée.*

3.6 Discussion

Dans ce chapitre, nous avons proposé deux procédures d'élicitation des experts pour intégrer des informations préliminaires dans le modèle Bliss. Les objectifs étaient d'avoir une procédure simple pour les experts et de pouvoir contrôler l'impact de ces connaissances d'experts sur l'inférence.

Approche 1 : pseudo-données La première approche se base sur la prise en compte de données construites par les experts, dites pseudo-données. La distribution *a priori* est

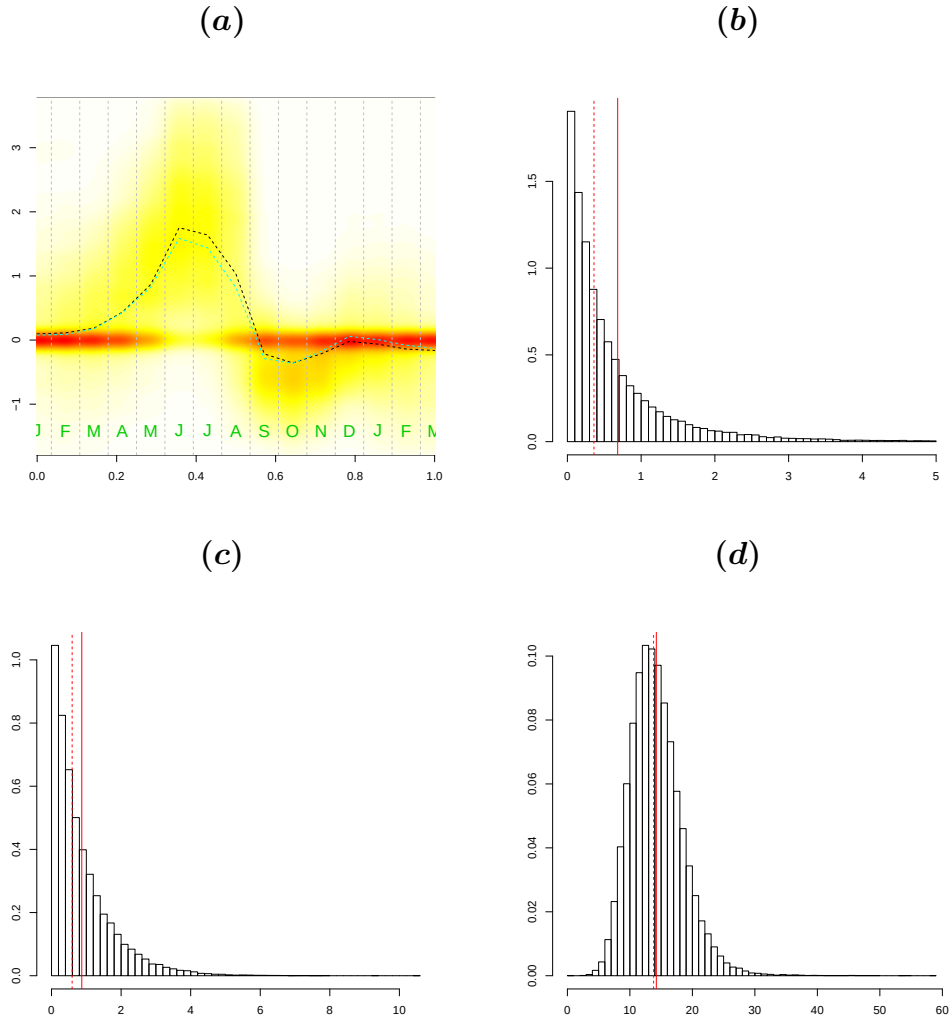


Figure 3.17: Résultats sur les données de truffes quand $\tau \sim \mathcal{E}(0)$. *Le graphique* (a) *(resp.* (b)*) représente la distribution a posteriori de la fonction coefficient (resp.* τ*). La courbe noire (resp. cyan) en pointillé est l'espérance a posteriori quand* $\tau \sim \mathcal{E}(0)$ *(resp. quand* τ *est fixé à 0 : pas de pénalisation). Le graphique* (c) *(resp.* (d)*) représente la distribution a posteriori du terme de pénalisation* $\tau \times \text{dist}^2(\beta^s, \bar{\beta}^s_E; \bar{g}_E)$ *(resp.* $\frac{1}{2\sigma^2} SCR$*). Pour les graphiques* (a)*,*(b) *et* (c)*, la droite verticale rouge en trait plein (resp. en pointillé) correspond à la moyenne (resp. médiane) de la distribution.*

la distribution du paramètre sachant les pseudo-données fournies par les experts.

Un premier avantage de cette approche est qu'il est possible de représenter la distribution *a priori* des experts en calculant la distribution du paramètre sachant les pseudo-données. En comparant cette distribution à la distribution *a posteriori* sachant les données observées et les pseudo-données, on peut visualiser l'apport des données observées et l'apport des pseudo-données.
Un autre avantage est le fait d'échanger avec les experts en discutant de quantités observables (données), et non pas de paramètres. Les pseudo-données qu'ils fournissent sont porteuses de leurs avis et leurs participations se résument à construire des scénarios, ce qui est intuitif pour eux.
De plus, nous avons introduit un poids par pseudo-données si bien que l'expert peut nuan-

cer ce qu'il pense pour chacune de ses données. Par exemple, un expert peut renseigner des scénarios qu'il considère réalistes mais il peut aussi renseigner des scénarios dont il n'est pas certain. Cela permet un certaine flexibilité pour les experts et une certaine finesse dans notre manière de modéliser leurs avis. Interagir avec les experts de cette manière permet de collecter assez efficacement leurs avis (Crowder, 1992; O'Hagan et al., 2006). En effet, à l'aide de la distribution *a priori* des experts, nous avons pu constater dans notre cas qu'effectivement les pseudo-données reflétaient correctement l'avis que les experts pouvaient formuler oralement.
La méthode permet également de contrôler l'impact sur l'inférence de la prise en compte d'avis d'experts au travers des poids w_i^e. On a notamment pu constater, analytiquement et numériquement, que lorsque les poids sont nuls, on peut considérer l'analyse comme objective.
Un autre aspect positif est que cette approche est basée sur une variante du modèle de régression qui n'induit pas de modification importante. Notamment en ce qui concerne les distributions conditionnelles complètes, voir l'annexe 3.7.1, même si cela nécessite d'utiliser un échantillonneur *a posteriori* un peu plus élaboré, voir l'annexe 3.7.2. Concernant la calibration des poids, nous avons suggéré une manière de procéder, mais il est possible d'utiliser une autre méthode de calibration sans que cela ne change grandement le reste de la modélisation.

Cependant, une des limitations de cette approche concerne l'interprétation de la distribution *a priori*. En effet, il n'est pas forcément évident de faire le lien entre la dispersion de la distribution et l'incertitude des experts. On aimerait pouvoir interpréter cette dispersion puisqu'elle a un impact certain sur la distribution *a posteriori*.
Une autre limitation est la difficulté pour les experts de renseigner des scénarios de précipitations. Faisant face à cette difficulté, nous avons suggéré nous-même des scénarios de précipitations aux experts et leur avons demander de fournir les productions associées.
De plus, il peut être difficile de fixer le coefficient de dépendance entre les experts. Nous les avons fixées nous-même par rapport à ce que nous considérons de cet ensemble d'experts, mais évidement cette méthode comporte de l'arbitraire et des biais.
Enfin, nous aurions pu prendre en compte plus finement ces dépendances. Par exemple, en corrigeant les poids de pseudo-données issues d'experts dépendants, en fonction de la proximité des scénarios de précipitations renseignés.

En perspective, il serait intéressant d'examiner comment cette méthode d'élicitation peut s'adapter à d'autres modèles. L'approche que nous avons proposée semble générique dans le sens où il suffit aux experts de renseigner des données et pour le statisticien de calibrer des poids. Suivant le contexte, il peut être plus ou moins direct d'obtenir des résultats équivalents aux propriétés 3.1 et 3.2.

Approche 2 : pénalisation Nous avons proposé une seconde approche basée sur une notion de pénalisation pour obtenir des distributions *a posteriori* qui sont *en accord* avec l'avis des experts.

Cette notion de pénalisation permet d'avoir une modélisation intuitive pour un statisticien et de mieux savoir comment interpréter les résultats. En effet, en faisant intervenir une notion de pénalisation, cela évoque des méthodes du type *Ridge* ou *Lasso*. En intro-

duisant une pénalité, ces méthodes ont pour objectif de favoriser des valeurs de coefficients de pente proches de (ou égales à) 0 pour réaliser un meilleur compromis biais/variance (ou pour sélectionner des variables). En transposant ces idées à notre contexte, nous comprenons que la pénalité introduite revient à favoriser des valeurs de θ qui sont en accord avec l'avis des experts. De plus, il est possible d'interpréter τ en termes d'intensité de pénalisation.
Un autre avantage est que dans le cadre de l'étude de l'impact des précipitations sur la production de truffes, cette approche a été efficace pour recueillir l'avis des experts sans que cela soit trop complexe pour eux. De plus, nous avons proposé deux méthodes pour traiter le problème du choix de l'hyperparamètre τ. La première consiste à fixer τ en utilisant une méthode bayésienne de validation croisée et la seconde considère une distribution *a priori* pour τ.

Cependant, suivant le contexte, les deux méthodes n'ont pas les mêmes performances. Pour ce qui est de la procédure de validation croisée bayésienne, il est nécessaire d'approcher un critère (l'utilité du modèle) ce qui donne deux sources de variabilité. La première vient du sous-échantillonnage sur les données et la seconde vient de l'approximation par échantillonnage préférentiel. Un des facteurs limitant de cette approche est l'utilisation du sous-échantillonnage lorsque le jeu de données est petit. Le sous-échantillon obtenu est de taille réduite et est utilisé pour calculer une approximation Monte Carlo. L'erreur d'approximation admet alors une variance importante (voir Vehtari and Ojanen, 2012). Or, les méthodes d'élicitation sont souvent utilisées lorsque le nombre de données est faible. Ce qui pose donc ici un problème à la calibration de τ en utilisant une procédure de validation-croisée.
D'un autre côté, si une loi *a priori* est considérée pour τ, nous avons observé avec les applications sur des données simulées et sur les données de truffes que la prise en compte de l'avis des experts n'a pas induit un changement notable dans la distribution *a posteriori*.
Enfin, cette approche est spécifique au modèle de régression linéaire fonctionnel. Comme souvent lorsque la distribution *a priori* est construite sur la base d'avis d'experts, certains choix autour de la modélisation sont dépendants du contexte d'application. Par exemple, nous avons construit un paramètre (la fonction signe) qui est "interprétable" dans le sens où il correspond à une caractéristique de la fonction coefficient intuitive pour les experts. Ainsi, l'approche que nous avons développée ici nécessite un travail non négligeable pour l'appliquer à d'autres contextes.

Une perspective de travail concerne la calibration de τ. Le choix de τ serait plus simple si nous pouvions déterminer un équivalent échantillon pour τ, comme par exemple pour le *g-prior* de Zellner (Zellner, 1986). De premier abord, il ne paraît pas évident de trouver un tel équivalent. En effet, les calculs font intervenir des quantités complexes comme par exemple les $x_i(\mathcal{I})$, qui sont doublement aléatoires (dépendent des données $x_i(t)$ et des hyperparamètres m et ℓ), ou encore la fonction signe $\beta^s(\cdot)$ (transformation complexe des hyperparamètres, voir (3.13)).

3.7 Annexe

3.7.1 Distributions conditionnelles complètes

Soient $\mathcal{D}$ des données observées et $\mathcal{D}_E$ des pseudo-données.

Nous donnons ci-dessous les distributions conditionnelles complètes dans le cadre du modèle décrit en section 3.2 :

$$\mu|\mathcal{D},\mathcal{D}_E,b,\sigma^2,m,\ell \sim \mathcal{N}\Big(\frac{\mathbf{1}_n^T(y-x_{\cdot}(\mathcal{I}_{\cdot})b)+\sum_{e=1}^E \mathbf{1}_{n_e}^T W^e(y^e-x_{\cdot}^e(\mathcal{I}_{\cdot})b)}{n+v_0^{-1}+\sum_{e=1}^E n_e w^e}, \frac{\sigma^2}{n+v_0^{-1}+\sum_{e=1}^E n_e w^e}\Big)$$

$$b|\mathcal{D},\mathcal{D}_E,\mu,\sigma^2,m,\ell \sim \mathcal{N}_K\Big(\hat{b}_1+\sum_{e=1}^E \hat{b}_{2,e}\,,\,\sigma^2 M_w^{-1}\Big)$$

$$\sigma^2|\mathcal{D},\mathcal{D}_E,\mu,b,m,\ell \sim \Gamma^{-1}\left(\frac{n+\sum_{e=1}^E n_e w^e+K+1}{2}, \frac{\mathrm{SCR}+\sum_{e=1}^E \mathrm{SCR}_e}{2}+\frac{1}{2}(\mu,b)^T\underline{\mathrm{V}}^{-1}(\mu,b)\right)$$

$$\pi\left(m_k|\mathcal{D},\mathcal{D}_E,\mu,b,\sigma^2,m_{-k},\ell\right) \propto \exp\left(-\frac{\mathrm{SCR}+\sum_{e=1}^E \mathrm{SCR}_e}{2\sigma^2}\right)\times\pi(b|m,\ell,\sigma^2)$$

$$\pi\left(\ell_k|\mathcal{D},\mathcal{D}_E,\mu,b,\sigma^2,m,\ell_{-k}\right) \propto \exp\left(-\frac{\mathrm{SCR}+\sum_{e=1}^E \mathrm{SCR}_e}{2\sigma^2}\right)\times\pi(\ell_k)\times\pi(b|m,\ell,\sigma^2)$$

où W^e est la matrice diagonale dont le i^e élément est w_i^e, $\mathrm{SCR}=\|y-\mu\mathbf{1}_n-x_{\cdot}(\mathcal{I}_{\cdot})b\|^2$, $\underline{\mathrm{V}}$ est donnée dans l'annexe 2.6.2 du chapitre 2 et

$$\mathrm{SCR}_e=(y^e-\mu\mathbf{1}_{n_e}-x_{\cdot}^e(\mathcal{I}_{\cdot})b)^T W^e(y^e-\mu\mathbf{1}_{n_e}-x_{\cdot}^e(\mathcal{I}_{\cdot})b).$$

Nous donnons ci-dessous les distributions conditionnelles complètes pour le modèle décrit en section 3.3 :

$$\mu|\mathcal{D},b,\sigma^2,m,\ell,\tau \sim \mathcal{N}\left(\frac{\mathbf{1}_n^T(y-x_{\cdot}(\mathcal{I}_{\cdot})b)}{n+v_0^{-1}}, \frac{\sigma^2}{n+v_0^{-1}}\right)$$

$$\pi(b|\mathcal{D},\mu,\sigma^2,m,\ell,\tau) \propto \exp\left\{-\left(\frac{\mathrm{SCR}+b^T\Sigma(\mathcal{I})^{-1}b}{2\sigma^2}+\tau\,\mathrm{dist}^2(\beta^s,\bar{\beta}_E^s;\bar{g}_E\right)\right\}$$

$$\pi(m|\mathcal{D},\mu,b,\sigma^2,\ell,\tau) \propto \exp\left\{-\left(\frac{\mathrm{SCR}}{2\sigma^2}+\tau\,\mathrm{dist}^2(\beta^s,\bar{\beta}_E^s;\bar{g}_E\right)\right\}$$

$$\pi(\ell|\mathcal{D},\mu,b,\sigma^2,m,\tau) \propto \exp\left\{-\left(\frac{\mathrm{SCR}}{2\sigma^2}+\tau\,\mathrm{dist}^2(\beta^s,\bar{\beta}_E^s;\bar{g}_E\right)\right\}\pi(\ell)$$

$$\pi(\tau|\mathcal{D},\mu,b,\sigma^2,m,\ell) \propto \exp\left\{-\tau\left(\lambda+\mathrm{dist}^2(\beta^s,\bar{\beta}_E^s;\bar{g}_E\right)\right\}.$$

La distribution de σ^2 est la même que celle donnée en annexe 2.6.2 dans le chapitre 2.

3.7.2 Implémentation

Nous présentons l'algorithme de *Metropolis-Within-Gibbs* que nous utilisons pour échantillonner la distribution *a posteriori* décrite en section 3.3.

Algorithme : *Metropolis-Within-Gibbs*

- Déterminer un point de départ pour θ_0 et τ.
- Déterminer la valeur de r.
  - $1-i$ Proposer une valeur pour r.
  - $1-ii$ Échantillonner la distribution *a posteriori* avec les étapes $3-i$ à $3-v$.
  - $1-iii$ Si le taux d'acceptation de l'étape Metropolis est entre 0.2 et 0.5, continuer sinon, aller à l'étape $1-i$ et proposer un nouveau r.
- Calculer τ_N grâce à l'échantillon *a posteriori* obtenu en $1-ii$ et déterminer le vecteur $\boldsymbol{\tau}$ allant de 0 à τ_N.
- Pour chaque valeur de $\boldsymbol{\tau}$:
  - $2-i$ Obtenir θ_t pour $t = 1, \ldots, T$ en exécutant les étapes $3-i$ à $3-v$.
  - $2-ii$ Calculer les termes $\pi(y_i|x_i, \theta_t; \tau)$ en utilisant un échantillonnage préférentiel décrit en section 3.3.2.
  - $2-iii$ En déduire l'utilité $u_{\text{IS-LOO}}(\tau)$.
- Choisir la valeur de τ parmi $\boldsymbol{\tau}$ qui admet la plus grande utilité $u_{\text{IS-LOO}}(\tau)$.
- Échantillonner la distribution *a posteriori*. Pour i de 1 à N:
  - $3-i$ Simuler $\mu_i \sim \mu|y, b_{i-1}, \sigma^2_{i-1}, m_{i-1}, \ell_{i-1}; \tau$,
  - $3-ii$ Simuler $\sigma^2_i \sim \sigma^2|y, \mu_i, b_{i-1}, m_{i-1}, \ell_{i-1}; \tau$,
  - $3-iii$ Simuler $m_i \sim m|y, \mu_i, b_{i-1}, \sigma^2_i, \ell_{i-1}; \tau$,
  - $3-iv$ Simuler $\ell_i \sim \ell|y, \mu_i, b_{i-1}, \sigma^2_i, m_i; \tau$,
  - $3-v$ Mettre à jour b en utilisant une étape de *Metropolis* :
    - $3-v-a$ Générer une proposition $b' \sim \mathcal{N}(b_{i-1}, r)$.
    - $3-v-b$ Calculer le taux d'acceptation α.
    - $3-v-c$ Simuler $u \sim \text{Unif}(0, 1)$.
    - $3-v-d$ Si $u < \alpha$ alors $b_i := b'$ sinon $b_{i-1} := b'$.

IV

Consistance de la méthode Bliss

Contents

4.1 Introduction 93
4.2 Notations 95
4.3 Hypothèses 96
4.4 Résultat 97
4.5 Preuves et lemmes 98
4.6 Discussions 105
4.7 Annexes 107

4.1 Introduction

Dans le chapitre 2, nous avons introduit le modèle Bliss, cas particulier du modèle de régression linéaire bayésien (2.3). Nous avons défini une loi *a priori* qui charge $\mathcal{E}_K$ l'ensemble des fonctions coefficient qui s'expriment comme des fonctions en escalier :

$$\beta(t) = \sum_{k=1}^{K} b_k \mathbf{1}_{\mathcal{I}_k}(t). \tag{4.1}$$

Par commodité, nous considérons que t appartient à $[0, 1]$. Le modèle Bliss pour $\theta = (\mu, \beta, \sigma^2)$, est donné par

$$y_i | x_i, \theta \overset{ind}{\sim} \mathcal{N}\left(r(x_i)\,,\, \sigma^2\right) = P_{\theta,i}, \quad \text{pour } i = 1, \ldots, n, \tag{4.2}$$

où $r(x_i) = \mu + \int_0^1 x_i(t)\beta(t)\,\mathrm{d}t$. La distribution *a priori* de θ est donnée par (2.9). Le modèle Bliss (4.2) dépend d'un choix de K et nous proposons en section 2.2.3 du chapitre 2 une manière de le choisir.

Supposons que la vraie fonction β_* est en escalier, avec K_* escaliers. Comme les intervalles peuvent se chevaucher, intuitivement il semble que si n est suffisamment grand, la loi *a posteriori* de β se concentre autour de β_* dès que $K_* \leq K$. Ainsi, le problème du choix de K se ramène en pratique à fixer un K suffisamment grand. Inversement, si $K_* > K$, on peut se demander si la loi *a posteriori* se concentre encore et si oui, autour de quel paramètre. Plus généralement, si la vraie fonction β_* n'est pas en escalier, ce qui est sûrement le cas en pratique, la distribution *a posteriori* se concentre-t-elle et autour de quelle valeur ? Ainsi, par la suite, on supposera que le modèle est mal spécifié et on considérera que les données sont générées selon :

$$y_i|x_i, \theta_* \overset{ind}{\sim} \mathcal{N}\left(r_*(x_i)\,,\, \sigma_*^2\right) = P_{\theta_*,i}, \qquad \text{pour } i = 1, \ldots, n, \tag{4.3}$$

où $r_*(x_i) = \mu_* + \int_0^1 x_i(t)\beta_*(t)\,\mathrm{d}t$, et avec $\beta_* \in L^2([0,1])$ qui n'est donc pas nécessairement une fonction en escalier.

Concernant l'étude du comportement asymptotique de la distribution *a posteriori*, Doob (1949) donne un premier résultat établissant la consistance pour tout θ n'étant pas dans un ensemble négligeable par rapport à la loi *a priori*. Ce premier théorème, utilisant un résultat de convergence des martingales, donne une convergence presque sûre et ne permet pas de vérifier la consistance en une valeur donnée de θ. Le théorème de Wald (1949) établit la consistance pour toute valeur dans le support de la loi *a priori* et montre la convergence de l'estimateur du maximum de vraisemblance. Pour décrire comment la distribution *a posteriori* se concentre, le théorème de Bernstein - von Mises donne une approximation de la loi *a posteriori* limite (voir Van der Vaart, 2000). Un des résultats les plus importants est le théorème de Schwartz (1965) qui établit la concentration de la distribution *a posteriori* sous les conditions 1) que la loi *a priori* charge un voisinage de Kullback-Leibler du vrai paramètre et 2) qu'il existe une suite consistante de fonctions de test. Par la suite, Ghosal et al. (2000) et Shen and Wasserman (2001) ont étendu ce résultat en donnant des vitesses de concentration. Pour un résumé des variantes de ces résultats et leurs déclinaisons dans différents contextes, on pourra consulter Ghosal (1997); Wasserman (1998); Van der Vaart (2000); Ghosh and Ramamoorthi (2003); Walker (2004); Castillo (2014); Kleijn (2016).

Deux variantes nous intéressent dans ce chapitre. La première concerne les résultats de consistance dans le cas où le modèle est mal spécifié. Les premiers résultats donnés par Berk et al. (1966); Huber (1967) se restreignent aux modèles paramétriques pour des données indépendantes et identiquement distribuées et ne s'appliquent donc pas dans notre contexte de régression non-paramétrique. Kleijn and van der Vaart (2006) propose un cadre général pour établir la consistance de modèles non-paramétriques mal spécifiés. Cependant, il nous a paru complexe de vérifier les hypothèses proposées dans notre contexte. Il nous semble plus efficace d'utiliser des outils moins sophistiqués pour montrer la consistance du modèle Bliss.
La seconde variante qui nous intéresse concerne l'étude des données non-identiquement distribuées. C'est le type de données que nous retrouvons pour des modèles de régression puisque la loi de y_i sachant x_i est différente de celle de y_j sachant x_j. Amewou-Atisso et al. (2003) proposent une version du théorème de Schwartz dans un modèle de régression semi-paramétrique et Choi and Schervish (2004); Xiang and Walker (2013) pour un modèle non-paramétrique. Pour obtenir des vitesses de concentration, Ghosal et al. (2007) généralisent le résultat de Ghosal et al. (2000) au cadre général de données qui ne sont

ni indépendantes, ni identiquement distribuées. Ces résultats et les travaux qui en ont découlé supposent généralement que le modèle est bien spécifié ou du moins qu'il l'est en partie (voir par exemple Choi, 2009 qui considère un modèle non-paramétrique avec des erreurs non-gaussiennes). De plus, ces travaux n'étudient pas le modèle de régression pour des données fonctionnelles. On notera comme exception le récent article de Lian et al. (2016) qui étudie la consistance d'un modèle de régression fonctionnelle bien spécifié.

Dans notre contexte, les données sont considérées indépendantes mais non-identiquement distribuées et nous travaillons avec un modèle de régression mal spécifié. Pour répondre à notre problématique tout en dégageant des hypothèses simples et interprétables, nous donnons un premier résultat : une version du théorème de Wald. Nous introduisons les notations en section 4.2. Les hypothèses du théorème sont détaillées en section 4.3. Le résultat principal et les grandes lignes de la démonstration sont données en section 4.4. En section 4.5, nous présentons le détails des preuves. Nous discutons du résultat et des extensions possibles en section 4.6.

4.2 Notations

On dispose d'un échantillon de n variables aléatoires réelles indépendantes y_i générées selon $P_{\theta_*,i}$ et d'une séquence de n courbes $x_i \in L^2([0,1])$ telles que :

$$\|x_i\|_{L^2}^2 = \int_0^1 x_i^2(t)\,\mathrm{d}t < \infty.$$

Les distributions $P_{\theta_*,i}$ et $P_{\theta,i}$ admettent des densités $p_{\theta_*,i}$ et $p_{\theta,i}$ par rapport à la mesure de Lebesgue pour tout $i \geq 1$ et tout $\theta \in \Theta$. Les distributions produits du modèle sont notées P_θ^n et P_θ^∞ et les distributions produits du *vrai* modèle sont notées $P_{\theta_*}^n$ et $P_{\theta_*}^\infty$. Pour une fonction f, l'espérance de f sous $P_{\theta_*,i}$ est notée $P_{\theta_*,i}f = \int f(y)\,P_{\theta_*,i}(\mathrm{d}y)$ et sous $P_{\theta_*}^\infty$ l'espérance est notée $P_{\theta_*}^\infty f = \int f(y)\,P_{\theta_*}^\infty(\mathrm{d}y)$. Le paramètre $\theta = (\mu, \beta, \sigma^2)$ appartient à $\Theta \subset \mathbb{R} \times \mathcal{E}_K \times \mathbb{R}_+^*$. Suivant le contexte, on utilisera la précision τ ou la variance σ^2. Pour travailler efficacement avec la vraisemblance

$$p_{\theta,i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - r(x_i)\right)^2\right\},$$

nous munissons Θ de la norme $\|\theta\| = |\tau| + \|\beta\|_{L^2} + |\mu|$.

La loi *a priori* sur Θ est notée par Π. Nous notons par Π_n la distribution *a posteriori* sur Θ sachant les données $\{y_i, x_i\}$ pour $i = 1, \ldots, n$. Le modèle étant dominé, nous travaillons avec (la version de) la distribution *a posteriori* donnée par

$$\Pi_n(U) = \frac{\int_U \prod_{i=1}^n p_{\theta,i}(y_i)\,\Pi(\mathrm{d}\theta)}{\int \prod_{i=1}^n p_{\theta,i}(y_i)\,\Pi(\mathrm{d}\theta)}, \tag{4.4}$$

pour tout ensemble U mesurable.

Pour f et g dans $L^2([0,1])$, on note par $f \otimes g$ l'élément h de $L^2([0,1]^2)$ qui vérifie $h(t,t') = f(t) \times g(t')$. On note aussi $f^{\otimes 2} = f \otimes f$ et $\|f^{\otimes 2}\|_{L^2}^2 = \iint f^{\otimes 2}(t,t')\,\mathrm{d}t\mathrm{d}t'$. De plus, on note par $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ la moyenne empirique de $x_1, \ldots, x_n$ et $\overline{x_n^2} = \frac{1}{n}\sum_{i=1}^n x_i^{\otimes 2}$.

4.3 Hypothèses

Contrairement aux cas où le modèle est bien spécifié, il est connu que la distribution *a posteriori* d'un modèle mal spécifié ne se concentre pas autour du *vrai* paramètre θ_*. Dans cette section, nous donnons des hypothèses afin d'établir que la distribution *a posteriori* du modèle Bliss se concentre autour d'un paramètre $\theta_0 \in \Theta$. En premier lieu, nous discutons des hypothèses sur les données fonctionnelles x_i.

Hypothèse 1. *Il existe $M < \infty$ tel que*

$$\sup_{\forall i \geq 1} \sup_{t \in [0,1]} |x_i(t)| \leq M.$$

L'hypothèse 1 est raisonnable dans un contexte d'application. Lorsqu'on mesure des grandeurs physiques telles que des précipitations ou des températures, on peut supposer qu'à tout instant t, les mesures $x_i(t)$ ne dépassent pas un certain seuil. Cette hypothèse implique que $\|x_i\|_{L^2} \leq M$, pour tout $i \geq 1$. Par soucis de simplicité, on suppose que $M = 1$.

Hypothèse 2. *Il existe $e \in L^2([0,1])$ et $c \in L^2([0,1]^2)$ tels que pour tout $t \in [0,1]$:*

$$|\bar{x}_n(t) - e(t)| \to 0 \qquad \left|\overline{x_n^2}(t,t') - c(t,t')\right| \to 0.$$

Avec l'hypothèse 2, on suppose que les courbes x_i admettent une certaine régularité ce qui est usuel lorsque le *design* est supposé déterministe. Cette hypothèse est analogue aux hypothèses qu'on retrouve dans d'autres contextes de régression. De plus l'hypothèse 1 implique que $\bar{x}_n$ converge vers e au sens de $L^2([0,1])$ et $\overline{x_n^2}$ vers c au sens de $L^2([0,1]^2)$. Si on considère un *design* aléatoire, cette hypothèse s'apparente à une loi des grands nombres sur les deux premiers moments et $c - e^{\otimes 2}$ se comprend comme une fonction de covariance.

Hypothèse 3. *Il existe une unique fonction $\beta_0 = \sum_{k=1}^{K} b_{0k} \mathbf{1}_{\mathcal{I}_{0k}} \in \mathcal{E}_K$ qui minimise*

$$F(\beta) = \iint (\beta_* - \beta)^{\otimes 2}(t,t') \left[c(t,t') - e^{\otimes 2}(t,t')\right] \mathrm{d}t \, \mathrm{d}t'. \tag{4.5}$$

Avec l'hypothèse 3, on suppose l'existence et l'unicité d'une solution β_0 sur le sous-ensemble $\mathcal{E}_K$. La fonction β_0 se comprend comme l'élément de $\mathcal{E}_K$ le plus *proche* de β_* au sens de F.

Hypothèse 4. *L'espace paramétrique $(\Theta, \|\cdot\|)$ est compact.*

Cette hypothèse forte permet d'avoir un cadre simplifié pour établir le résultat. En particulier, elle implique qu'il existe un $\eta < \infty$ tel que

$$\sup_{\theta \in \Theta} \|\theta\| \leq \eta.$$

4.4 Résultat

Nous commençons avec la proposition 4.1 par caractériser θ_0 à partir du *design* et de θ_*. De plus, cette proposition montre que θ_0 est le paramètre le plus *proche* de θ_* au sens de la divergence de Kullback-Leibler, ce qui est généralement ce qu'on obtient dans d'autres contextes.

Proposition 4.1. *Sous les hypothèses 1 à 4, la fonction*

$$\theta \mapsto D_{KL}\left(P_{\theta_*}^{\infty}, P_{\theta}^{\infty}\right) = P_{\theta_*}^{\infty} \log \frac{p_{\theta_*}^{\infty}}{p_{\theta}^{\infty}} \tag{4.6}$$

admet un unique minimum sur Θ *en* $\theta_0 = (\mu_0, \beta_0, \sigma_0^2)$ *où* β_0 *est donné par l'hypothèse 3,* $\mu_0 = \mu_* + \int (\beta_* - \beta_0)(t)\, e(t)\, \mathrm{d}t$ *et* $\sigma_0^2 = \sigma_*^2 + F(\beta_0)$.

Dans un premier temps, on remarque que si β_* est une fonction en escalier, avec K escaliers ou moins, alors θ_0 coïncide avec le *vrai* paramètre $(\mu_*, \beta_*, \sigma_*^2)$. Plus généralement, la divergence de Kullback-Leibler (4.6) atteint son minimum en particulier si F atteint un minimum en β. À partir des hypothèses 1 et 2 sur le *design*, on montre avec la proposition 4.2 que $\Sigma(t,t) = c(t,t') - e^{\otimes 2}(t,t')$ est semi-définie positive sur $L^2([0,1])$. On montre alors que F est convexe ce qui implique qu'une solution existe sur $L^2([0,1])$.

Proposition 4.2. *Sous les hypothèses 1 et 2, la fonction* $\Sigma(t,t') = c(t,t') - e^{\otimes 2}(t,t')$ *est semi-définie positive : pour tout* $f \in L^2([0,1])$,

$$S(f) = \iint f^{\otimes 2}(t,t')\, \Sigma(t,t')\, \mathrm{d}t\mathrm{d}t' \geq 0.$$

De plus, $F(\beta) = S(\beta_* - \beta)$ *est convexe sur* $L^2([0,1])$.

Cependant, comme nous nous restreignons aux fonctions en escalier, nous aurons besoin de considérer le minimum de F sur le sous ensemble de fonctions $\mathcal{E}_K$. Avec l'hypothèse 3, nous supposons qu'il existe une unique solution de F sur $\mathcal{E}_K$. On obtient alors la consistance de la distribution *a posteriori* en $\theta_0 \in \Theta$ avec le théorème 4.3.

Théorème 4.3. *Soit* U *le complémentaire d'un voisinage de* θ_0. *Sous les hypothèses 1 à 4,*

$$\Pi_n(U) \to 0$$

$P_{\theta_*}^{\infty}$*-presque sûrement, quand* $n \to +\infty$.

Pour montrer le théorème, on fait apparaître au numérateur et au dénominateur de la distribution *a posteriori*, le terme $n^{-1} \sum_{i=1}^{n} - \log p_{\theta,i}/p_{\theta_0,i}$ qui par la loi forte des grands nombres tends vers

$$-P_{\theta_*}^{\infty} \log \frac{p_{\theta}^{\infty}}{p_{\theta_0}^{\infty}} \triangleq \bar{K}(\theta),$$

qui s'apparente à la divergence de Kullback-Leibler (4.6). Grâce à la proposition 4.1, on détermine que $\bar{K}$ est minimale en θ_0. Puis, on détermine un voisinage de θ_0 tel que $\bar{K}(\theta)$ soit plus petite à l'intérieur qu'à l'extérieur du voisinage. Ceci permet de contrôler le numérateur et le dénominateur pour en déduire la convergence de la distribution *a posteriori*. Il est aussi nécessaire de montrer que la distribution *a priori* charge un voisinage de θ_0.

4.5 Preuves et lemmes

Preuve de la proposition 4.1

Minimiser la divergence $D_{\mathrm{KL}}\left(P_{\theta_*}^{\infty}, P_{\theta}^{\infty}\right)$ revient à minimiser $\bar{Q}(\theta) = -P_{\theta_*}^{\infty} \log p_{\theta}^{\infty}$. Par le lemme 4.4, on a l'expression

$$\bar{Q}(\theta) = \bar{Q}(\mu, \beta, \sigma^2) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\Big[\sigma_*^2 + G(\mu, \beta) + F(\beta)\Big],$$

où F est donnée par l'hypothèse 3 et $G(\mu, \beta) = \left(\mu_* - \mu + \int(\beta_* - \beta)(t)e(t)\,\mathrm{d}t\right)^2$. Pour minimiser $\bar{Q}(\theta)$, on minimise de manière séquentielle en μ, en σ^2 puis en β. On obtient en minimisant sur μ que $\bar{Q}(\mu, \beta, \sigma^2) \geq \bar{Q}(\mu_0, \beta, \sigma^2)$ où

$$\mu_0 = \mu_* + \int(\beta_* - \beta)(t)e(t)\,\mathrm{d}t,$$

ce qui implique que pour tout β, $G(\mu_0, \beta) = 0$. Pour minimiser en σ^2, on étudie la dérivée

$$\frac{\partial \bar{Q}(\mu_0, \beta, \sigma^2)}{\partial \sigma^2} = \frac{\sigma^2 - (\sigma_*^2 + F(\beta))}{2\sigma^4}.$$

Puisque $F(\beta) \geq 0$ pour tout β par la proposition 4.2, $\bar{Q}(\mu_0, \beta, \sigma^2)$ admet un minimum en $\sigma_0^2 = \sigma_*^2 + F(\beta)$. On a donc

$$\begin{aligned}\bar{Q}(\mu_0, \beta, \sigma_0^2) &= \frac{1}{2}\log 2\pi\sigma_0^2 + \frac{1}{2\sigma_0^2}\left[\sigma_*^2 + F(\beta)\right] \\ &= \frac{1}{2}\log 2\pi e[\sigma_*^2 + F(\beta)].\end{aligned}$$

Donc si β_0 minimise F, alors $\bar{Q}(\mu_0, \beta, \sigma_0^2) \geq \bar{Q}(\mu_0, \beta_0, \sigma_0^2)$. D'où pour tout $\theta \in \Theta$, $\bar{Q}(\mu, \beta, \sigma^2) \geq \bar{Q}(\mu_0, \beta_0, \sigma_0^2)$ où

$$\mu_0 = \mu_* + \int(\beta_* - \beta_0)(t)e(t)\,\mathrm{d}t, \quad \sigma_0^2 = \sigma_*^2 + F(\beta_0)$$

et β_0 est donné par l'hypothèse 3. □

Preuve de la proposition 4.2

Soient $d_n = n^{-1}\sum x_i^{\otimes 2} - (n^{-1}\sum x_i)^{\otimes 2} = \overline{x_n^2} - \bar{x}_n^{\otimes 2}$ et $S_n(f) = \iint f^{\otimes 2}(t, t')\, d_n(t, t')\,\mathrm{d}t\mathrm{d}t'$. D'après l'hypothèse 2, la suite $(d_n)_n$ converge simplement vers $c - e^{\otimes 2}$. De plus, on peut

réécrire d_n à la manière du théorème de König-Huygens :

$$\begin{aligned}
d_n(t,t') &= \overline{x_n^2}(t,t') - \bar{x}_n^{\otimes 2}(t,t') \\
&= \overline{x_n^2}(t,t') + \bar{x}_n^{\otimes 2}(t,t') - n^{-1}\sum_{i=1}^n x_i(t)\bar{x}_n(t') - n^{-1}\sum_{i=1}^n x_i(t')\bar{x}_n(t) \\
&= n^{-1}\sum_{i=1}^n x_i(t)x_i(t') + n^{-1}\sum_{i=1}^n \bar{x}_n^{\otimes 2}(t,t') - n^{-1}\sum_{i=1}^n x_i(t)\bar{x}_n(t') - n^{-1}\sum_{i=1}^n x_i(t')\bar{x}_n(t) \\
&= n^{-1}\sum_{i=1}^n \left[x_i(t)x_i(t') + \bar{x}_n(t)\bar{x}_n(t') - x_i(t)\bar{x}_n(t') - x_i(t')\bar{x}_n(t) \right] \\
&= n^{-1}\sum_{i=1}^n \left(x_i - \bar{x}_n\right)^{\otimes 2}(t,t').
\end{aligned}$$

D'où, pour tout $f \in L^2([0,1])$ et tout $n \geq 1$,

$$S_n(f) = n^{-1}\sum_{i=1}^n \left(\int f(t)\left(x_i - \bar{x}_n\right)(t)\,\mathrm{d}t \right)^2 \geq 0.$$

Avec l'hypothèse 1, on peut dominer $f^{\otimes 2} \times d_n$:

$$\begin{aligned}
|f(t)f(t')d_n(t,t')| &\leq n^{-1}\sum_{i=1}^n |f(t)f(t')| \left[|x_i(t)| + n^{-1}\sum_{j=1}^n |x_j(t)| \right] \left[|x_i(t')| + n^{-1}\sum_{j=1}^n |x_j(t')| \right] \\
&\leq 4\,|f(t)f(t')| \quad \triangleq g(t,t').
\end{aligned}$$

De plus, puisque $f \in L^2([0,1])$, la fonction dominante g est intégrable :

$$\iint g(t,t')\,\mathrm{d}t\mathrm{d}t' = 4\left(\int |f(t)|\,\mathrm{d}t \right)^2 \leq 4\|f\|_{L^2}^2.$$

Ainsi, par le théorème de convergence dominée, $S(f) = \lim S_n(f) \geq 0$ et donc $c - e^{\otimes 2}$ est semi-définie positive.

Soient β_1 et $\beta_2 \in L^2([0,1])$ et $\lambda \in [0,1]$. On montre la convexité de F par :

$$\begin{aligned}
F\Big(\lambda\beta_1 + (1-\lambda)\beta_2\Big) &= \iint \Big(\lambda\beta_* - \lambda\beta_1 + (1-\lambda)\beta_* - (1-\lambda)\beta_2\Big)^{\otimes 2}(t,t')\,\Sigma(t,t')\,\mathrm{d}t\mathrm{d}t' \\
&= \lambda F(\beta_1) + (1-\lambda)F(\beta_2) - \lambda(1-\lambda)S(\beta_1 - \beta_2) \\
&\leq \lambda F(\beta_1) + (1-\lambda)F(\beta_2).
\end{aligned}$$

Preuve du théorème 4.3

Soit $\theta_0 = (\mu_0, \beta_0, \sigma_0^2)$ donné par la proposition 4.1. On réécrit la distribution *a posteriori* (4.4) en faisant intervenir le logarithme du rapport de vraisemblance $p_{\theta,i}/p_{\theta_0,i}$:

$$\Pi_n(U) = \frac{\int_U \exp\left\{-n\left(n^{-1}\sum_{i=1}^n - \log p_{\theta,i}(y_i)/p_{\theta_0,i}(y_i)\right)\right\}\,\Pi(\mathrm{d}\theta)}{\int \exp\left\{-n\left(n^{-1}\sum_{i=1}^n - \log p_{\theta,i}(y_i)/p_{\theta_0,i}(y_i)\right)\right\}\,\Pi(\mathrm{d}\theta)}.$$

Quand il n'y aura pas d'ambiguïté, on remplacera $\sum_{i=1}^n$ par $\sum$. Le but est de majorer le rapport par un terme qui tend vers 0 lorsque n tend vers ∞. on reconnaît que$n^{-1}\sum_{i=1}^n - \log p_{\theta,i}(y_i)/p_{\theta_0,i}(y_i)$ tend vers un terme qui s'apparente à la divergence de Kullback-Leibler (4.6), minimale en θ_0 par la proposition 4.1. On pose

- $T_i(\theta, y_i) = -\log\left(p_{\theta,i}/p_{\theta_0,i}\right)$,
- $Q_i(\theta) = -P_{\theta_*,i}\,\log p_{\theta,i}$,
- $K_i(\theta) = P_{\theta_*,i}\,T_i(\theta) = Q_i(\theta) - Q_i(\theta_0)$ et
- $W_i(\theta, y_i) = -\log p_{\theta,i} - Q_i(\theta)$.

En remarquant que $T_i(\theta, y_i) - K_i(\theta) = W_i(\theta, y_i) - W_i(\theta_0, y_i)$, on écrit :

$$\begin{aligned}\frac{1}{n}\sum -\log\frac{p_{\theta,i}}{p_{\theta_0,i}}(y_i) &= \frac{1}{n}\sum\left(T_i(\theta, y_i) - K_i(\theta)\right) + \frac{1}{n}\sum K_i(\theta)\\ &= \frac{1}{n}\sum W(\theta, y_i) - \frac{1}{n}\sum W(\theta_0, y_i) + \frac{1}{n}\sum K_i(\theta) - \bar{K}(\theta) + \bar{K}(\theta)\\ &\triangleq R_n(\theta, y) + \bar{K}(\theta),\end{aligned}$$

où $\bar{K}(\theta) = \bar{Q}(\theta) - \bar{Q}(\theta_0)$. Par le lemme 4.4, les fonctions $\bar{K}$ et $\bar{Q}$ sont continues pour $\theta \in \Theta$ et par la proposition 4.1, $\bar{Q}$ admet un minimum unique en $\theta_0 \in \Theta$. Il suit que $\bar{K}$ admet un minimum unique en $\theta_0 \in \Theta$. Puisque sur le compact Θ, $\bar{K}$ est continue et admet un minimum unique, alors pour tout $\varepsilon > 0$ suffisamment petit et pour tout $\varepsilon' > \varepsilon$,

$$S \triangleq \sup_{\|\theta_0 - \theta\| \le \varepsilon} \bar{K}(\theta) < \inf_{\|\theta_0 - \theta\| \ge \varepsilon'} \bar{K}(\theta) \triangleq I. \tag{4.7}$$

On choisit alors ε et ε' de sorte que $U \subset \{\theta : \|\theta_0 - \theta\| \ge \varepsilon'\}$ et tel que (4.7) soit vérifiée. On pose $\delta = I - S$. Par le lemme 4.5, $\sup_{\theta\in\Theta} R_n(\theta, y) \to 0$, $P_{\theta_*}^{\infty}$-presque sûrement quand $n \to \infty$. Donc $P_{\theta_*}^{\infty}$-presque sûrement, il existe un n suffisamment grand tel que pour tout $\theta \in \Theta$,

$$\bar{K}(\theta) - \frac{\delta}{2} \le \frac{1}{n}\sum -\log\frac{p_{\theta,i}}{p_{\theta_0,i}}(y_i) \le \bar{K}(\theta) + \frac{\delta}{2}.$$

D'où, nous obtenons :

$$\sup_{\|\theta_0 - \theta\| \le \varepsilon} \frac{1}{n}\sum -\log\frac{p_{\theta,i}}{p_{\theta_0,i}}(y_i) \le S + \frac{\delta}{2} \qquad \text{et} \qquad I - \frac{\delta}{2} \le \inf_{\|\theta_0 - \theta\| \ge \varepsilon'} \frac{1}{n}\sum -\log\frac{p_{\theta,i}}{p_{\theta_0,i}}(y_i).$$

Donc, $P_{\theta_*}^{\infty}$-presque sûrement pour n suffisamment grand, la distribution *a posteriori* en U est majorée par :

$$\begin{aligned}\Pi_n(U) &\le \frac{\int_U \exp\left\{-n \inf\limits_{\|\theta_0 - \theta\| \ge \varepsilon'} \frac{1}{n}\sum \log\frac{p_{\theta,i}}{p_{\theta_0,i}}(y_i)\right\}\,\Pi(\mathrm{d}\theta)}{\int_{\|\theta_0 - \theta\| \le \varepsilon} \exp\left\{-n \sup\limits_{\|\theta_0 - \theta\| \le \varepsilon} \frac{1}{n}\sum \log\frac{p_{\theta,i}}{p_{\theta_0,i}}(y_i)\right\}\,\Pi(\mathrm{d}\theta)}\\ &\le \frac{\int_U e^{-n(I - \delta/2)}\,\Pi(\mathrm{d}\theta)}{\int_{\|\theta_0 - \theta\| \le \varepsilon} e^{-n(S + \delta/2)}\,\Pi(\mathrm{d}\theta)}\\ &= \frac{\Pi(U)}{\Pi(\|\theta_0 - \theta\| \le \varepsilon)}\,e^{-n(I - S)}\end{aligned}$$

Par le lemme 4.8, $\Pi(\|\theta_0 - \theta\| \le \varepsilon) > 0$, d'où on obtient le résultat. □

Lemme 4.4. *Sous les hypothèses 1, 2 et 4,* $n^{-1}\sum_{i=1}^n Q_i(\theta) = -n^{-1}\sum_{i=1}^n P_{\theta_*,i}\log p_{\theta,i}$ *converge uniformément vers* $\bar{Q}$ *pour* $\theta\in\Theta$ *où*

$$\bar{Q}(\theta) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\Bigg[\sigma_*^2 + \Big(\mu_* - \mu + \int(\beta_*-\beta)(t)\,e(t)\,\mathrm{d}t\Big)^2 + \iint(\beta_*-\beta)^{\otimes 2}(t,t')\,\Big[c(t,t') - e^{\otimes 2}(t,t')\Big]\,\mathrm{d}t\mathrm{d}t'\Bigg]$$

Démonstration. Comme $P_{\theta,i} = \mathcal{N}(r(x_i),\sigma^2)$ et $y_i\sim P_{\theta_*,i} = \mathcal{N}\Big(r_*(x_i),\sigma_*^2\Big)$, on a

$$\begin{aligned}-P_{\theta_*,i}\log p_{\theta,i} &= \frac{1}{2}\log 2\pi\sigma^2\int p_{\theta_*,i}(y_i)\,\mathrm{d}y_i + \frac{1}{2\sigma^2}\int\Big(y_i - r(x_i)\Big)^2 p_{\theta_*,i}(y_i)\,\mathrm{d}y_i\\ &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\Big[\sigma_*^2 + \Big(r_*(x_i)-r(x_i)\Big)^2\Big].\end{aligned}$$

Pour établir la convergence uniforme, on montre que la distance entre $\bar{Q}$ et la série au rang n est majorée par un terme qui ne dépend pas de θ et qui tend vers 0.

$$\Big|n^{-1}\sum Q_i(\theta) - \bar{Q}(\theta)\Big| = \frac{1}{2\sigma^2}\Big|n^{-1}\sum\Big(r_*(x_i)-r(x_i)\Big)^2 - (\mu_*-\mu)^2 - 2(\mu_*-\mu)\int(\beta_*-\beta)(t)e(t)\,\mathrm{d}t - \iint(\beta_*-\beta)^{\otimes 2}(t,t')c(t,t')\,\mathrm{d}t\mathrm{d}t'\Big|.$$

Comme

$$(r_*(x_i)-r(x_i))^2 = (\mu_*-\mu)^2 + 2(\mu_*-\mu)\int(\beta_*-\beta)(t)x_i(t)\,\mathrm{d}t + \iint(\beta_*-\beta)^{\otimes 2}(t,t')\,x_i^{\otimes 2}(t,t')\,\mathrm{d}t\mathrm{d}t',$$

on a :

$$\begin{aligned}\Big|n^{-1}\sum Q_i(\theta) - \bar{Q}(\theta)\Big| &= \frac{1}{2\sigma^2}\Big|2(\mu_*-\mu)\int(\beta_*-\beta)(t)\Big[n^{-1}\sum x_i(t) - e(t)\Big]\,\mathrm{d}t +\\ &\qquad \iint(\beta_*-\beta)^{\otimes 2}(t,t')\Big[n^{-1}\sum x_i^{\otimes 2}(t,t') - c(t,t')\Big]\,\mathrm{d}t\mathrm{d}t'\Big|\\ \text{(par Cauchy-Schwartz)} &\le \frac{1}{2\sigma^2}\Big[2|\mu_*-\mu|\times\|\beta_*-\beta\|_{L^2}\times\Big\|n^{-1}\sum x_i - e\Big\|_{L^2}\\ &\qquad + \|\beta_*-\beta\|_{L^2}^2\,\Big\|n^{-1}\sum x_i^{\otimes 2} - c\Big\|_{L^2}\Big]\end{aligned}$$

Comme Θ est compact, il existe un δ_* qui majore $\|\theta-\theta_*\|$ pour tout $\theta\in\Theta$. L'hypothèse 4 implique qu'il existe $\eta>0$ tel que $1/\sigma^2 = \tau < \eta$ pour tout $\theta\in\Theta$, d'où

$$\Big|n^{-1}\sum Q_i(\theta) - \bar{Q}(\theta)\Big| \le \frac{\eta\delta_*^2}{2}\Big[2\Big\|n^{-1}\sum x_i - e\Big\|_{L^2} + \Big\|n^{-1}\sum x_i^{\otimes 2} - c\Big\|_{L^2}\Big].$$

On conclut avec les hypothèses 1 et 2 qui impliquent les convergences L^2 de $n^{-1}\sum x_i$ vers e et de $n^{-1}\sum x_i^{\otimes 2}$ vers c. □

Lemme 4.5. *Sous les hypothèses 1, 2 et 4,*

$$\sup_{\theta\in\Theta} R_n(\theta,y)\to 0$$

$P_{\theta_*}^{\infty}$*-presque sûrement quand* $n\to\infty$.

Démonstration. Rappelons que $R_n(\theta, y) = n^{-1}\sum W_i(\theta, y_i) - n^{-1}\sum W_i(\theta_0, y_i) + n^{-1}\sum K_i(\theta) - \bar{K}(\theta)$ où

- $W_i(\theta, y_i) = -\log p_{\theta,i}(y_i) - Q_i(\theta)$
- $Q_i(\theta) = -P_{\theta_*,i}\log p_{\theta,i}$
- $K_i(\theta) = Q_i(\theta) - Q_i(\theta_0)$ et $\bar{K}(\theta) = \bar{Q}(\theta) - \bar{Q}(\theta_0)$.

On commence par montrer la convergence de $n^{-1}\sum K_i(\theta) - \bar{K}(\theta)$, qui ne dépend pas des y_i. Par le lemme 4.4, $n^{-1}\sum Q_i$ converge uniformément vers $\bar{Q}$ pour tout $\theta \in \Theta$. Comme $n^{-1}\sum K_i(\theta) - \bar{K}(\theta) = n^{-1}\sum Q_i(\theta) - \bar{Q}(\theta) + n^{-1}\sum Q_i(\theta_0) - \bar{Q}(\theta_0)$ et que $\theta_0 \in \Theta$, on a

$$\sup_{\theta\in\Theta} n^{-1}\sum K_i(\theta) - \bar{K}(\theta) \to 0 \text{ quand } n \text{ tend vers } \infty. \tag{4.8}$$

On montre maintenant la convergence $P_{\theta_*}^{\infty}$-presque sûre de $\sup_{\theta\in\Theta} n^{-1}\sum W_i(\theta, y_i)$. On note $\Delta_i(r) = \big(r_*(x_i) - r(x_i)\big)/\sigma_*$ et $Z_i = \big(y_i - r_*(x_i)\big)/\sigma_*$. On réécrit Q_i et W_i comme :

$$\begin{aligned}
\bullet\ Q_i(\theta) &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\left[\sigma_*^2 + \big(r_*(x_i) - r(x_i)\big)^2\right] \\
&= \frac{1}{2}\log 2\pi\sigma^2 + \frac{\sigma_*^2}{2\sigma^2}\big[1 + \Delta_i(r)^2\big] \\
\bullet\ W_i(\theta, y_i) &= -\log p_{\theta,i}(y_i) - Q_i(\theta) = \frac{1}{2\sigma^2}\left[\big(y_i - r(x_i)\big)^2 - \sigma_*^2 - \sigma_*^2\Delta_i(r)^2\right] \\
&= \frac{\sigma_*^2}{2\sigma^2}\left[Z_i^2 + 2Z_i\Delta_i(r) - 1\right]
\end{aligned}$$

Pour étudier $\sup_{\theta\in\Theta} n^{-1}\sum W_i(\theta, y_i)$, on utilise la compacité de Θ. Soient $\varepsilon > 0$ et $\delta > 0$ tels que $\delta c < \varepsilon$, où c est donné par le lemme 4.6. Soient $B_1, \ldots, B_J$ où

$$B_j = \{\theta \in \Theta : \|\theta - \theta_j\| \le \delta\} \quad \text{où } \theta_j \in \Theta.$$

tels que l'union des B_j recouvrent Θ. En notant $F_i(\theta_j, \delta) = \sup_{\theta\in B_j} |W_i(\theta, y_i) - W_i(\theta_j, y_i)|$, nous avons pour $\theta \in B_j$:

$$\begin{aligned}
\left|n^{-1}\sum W_i(\theta, y_i)\right| &\le n^{-1}\sum \sup_{\theta\in B_j} |W_i(\theta, y_i) - W_i(\theta_j, y_i)| + \left|n^{-1}\sum W_i(\theta_j, y_i)\right| \\
&\le \underbrace{n^{-1}\sum \left[F_i(\theta_j, \delta) - P_{\theta_*,i}\, F_i(\theta_j, \delta)\right]}_{\triangleq I_j} + \underbrace{n^{-1}\sum P_{\theta_*,i}\, F_i(\theta_j, \delta)}_{\triangleq II_j} + \underbrace{\left|n^{-1}\sum W(\theta_j, y_i)\right|}_{\triangleq III_j}.
\end{aligned}$$

Par le lemme 4.6, $n^{-1}\sum P_{\theta_*,i}\, F_i(\theta_j, \delta) < \delta c$. D'où, II_j (qui ne dépend pas des y_i) est majoré par ε. De plus, avec le lemme 4.6, il existe d indépendant de i et j tel que $P_{\theta_*,i}\, F_i(\theta_j, \delta)^2 \le \delta^2 d$, d'où

$$\sum_{i=1}^{\infty} \frac{P_{\theta_*,i}\, F_i(\theta_j, \delta)^2}{i^2} \le \delta^2 d \sum_{i=1}^{\infty} i^{-2} < \infty.$$

Donc, par la loi forte des grands nombres de Kolmogorov pour des variables aléatoires indépendantes et non identiquement distribuées (Sen and Singer, 1994, page 67), on obtient $I_j \to 0$, $P_{\theta_*}^{\infty}$-presque sûrement. Concernant III_j, comme $P_{\theta_*,i}\, W_i(\theta_j, y_i) = 0$ et que

pour un certain a, $P_{\theta_*,i}\, W_i(\theta_j, y_i)^2 < a$ par le lemme 4.7, alors par le même théorème de Kolmogorov, on obtient que $III_j \to 0$, $P_{\theta_*}^{\infty}$-presque sûrement. On a alors :

$$\begin{aligned}\sup_{\theta\in\Theta} n^{-1}\sum W_i(\theta, y_i) &\le \max_{j\le J}\sup_{\theta\in B_j}\left|n^{-1}\sum W_i(\theta, y_i)\right| \\ &\le \max_j I_j + \max_j II_j + \max_j III_j \\ &\le \sum_{j=1}^{J} I_j + \varepsilon + \sum_{j=1}^{J} III_j.\end{aligned}$$

Par la convergence $P_{\theta_*}^{\infty}$-presque sûre de I_j et III_j vers 0, il en résulte que

$$\sup_{\theta\in\Theta} n^{-1}\sum W_i(\theta, y_i) \to 0, \quad P_{\theta_*}^{\infty}\text{-presque sûrement.} \tag{4.9}$$

Comme $\theta_0 \in \Theta$, le résultat se déduit de (4.8) et (4.9). □

Lemme 4.6. *Sous l'hypothèse 4, il existe c et d tels que pour tout $\theta_1 \in \Theta$, $\delta > 0$ et $i \ge 1$,*

$$P_{\theta_*,i}\, F_i(\theta_1, \delta) \le \delta c \quad \text{et} \quad P_{\theta_*,i}\, F_i(\theta_1, \delta)^2 \le \delta^2 d.$$

Démonstration. On rappelle que $F_i(\theta_1, \delta) = \sup_{\theta\in B_1} |W_i(\theta, y_i) - W_i(\theta_1, y_i)|$. Soient $\delta > 0$ et $\theta \in \Theta$ tels que $\|\theta - \theta_1\| \le \delta$. On note $r_1(x_i) = \mu_1 + \int \beta_1(t)x_i(t)\,\mathrm{d}t$ et rappelons que $\Delta_i(r) = \big(r_*(x_i) - r(x_i)\big)/\sigma_*$ et $Z_i = \big(y_i - r_*(x_i)\big)/\sigma_*$, alors on peut écrire :

$$\begin{aligned}|W_i(\theta, y_i) - W_i(\theta_1, y_i)| &= \sigma_*^2\left|\frac{1}{2\sigma^2}\big[Z_i^2 + 2Z_i\Delta_i(r) - 1\big] - \frac{1}{2\sigma_1^2}\big[Z_i^2 + 2Z_i\Delta_i(r_1) - 1\big]\right| \\ &= \frac{\sigma_*^2}{2}\left|Z_i^2(\sigma^{-2} - \sigma_1^{-2}) + 2Z_i\left(\frac{\Delta_i(r)}{\sigma^2} - \frac{\Delta_i(r_1)}{\sigma_1^2}\right) - (\sigma^{-2} - \sigma_1^{-2})\right|.\end{aligned}$$

L'hypothèse 1 implique que $r(x_i) - r_1(x_i) = \mu - \mu_1, + \int(\beta_1 - \beta)(t)x_i(t)\,\mathrm{d}t \le \|\theta - \theta_1\|$. D'où,

$$\begin{aligned}|W_i(\theta, y_i) - W_i(\theta_1, y_i)| &\le \frac{\sigma_*^2}{2}\left|Z_i^2 + 2Z_i\frac{r_*(x_i)}{\sigma_*} - 1\right| \times |\sigma^{-2} - \sigma_1^{-2}| + \sigma_*^2\left|Z_i\sigma_*^{-1}\left(\frac{r(x_i)}{\sigma^2} - \frac{r_1(x_i)}{\sigma_1^2}\right)\right| \\ &\le \|\theta - \theta_1\|\frac{\sigma_*^2}{2}\left|Z_i^2 + 2Z_i\frac{r_*(x_i)}{\sigma_*} - 1\right| + \sigma_*\left|Z_i\left(\frac{r(x_i)}{\sigma^2} - \frac{r_1(x_i)}{\sigma_1^2}\right)\right|.\end{aligned}$$

En remarquant que $\sigma^{-2}r(x_i) - \sigma_1^{-2}r_1(x_i) = \frac{r(x_i)-r_1(x_i)}{\sigma_1^2} + r(x_i)\big(\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2}\big) \le \|\theta - \theta_1\| \times \big(\|\theta_1\| + \|\theta\|\big) \le 2\eta\|\theta - \theta_1\|$, on obtient :

$$\begin{aligned}|W_i(\theta, y_i) - W_i(\theta_1, y_i)| &\le \|\theta - \theta_1\|\left[\frac{\sigma_*^2}{2}\left|Z_i^2 + 2Z_i\frac{r_*(x_i)}{\sigma_*} - 1\right| + 2\eta\sigma_*|Z_i|\right] \\ &\triangleq \|\theta - \theta_1\|c_i.\end{aligned}$$

Comme c_i est indépendant θ, on a alors $0 \le F_i(\theta_1, \delta) \le \delta c_i$ et donc $P_{\theta_*,i}\, F_i(\theta_1, \delta) \le$

$\delta \times P_{\theta_*,i}\, c_i$. On obtient une majoration indépendante de i avec :

$$\begin{aligned} P_{\theta_*,i}\, c_i &= \frac{\sigma_*^2}{2} P_{\theta_*,i} \left| Z_i^2 + 2 Z_i \frac{r_*}{\sigma_*} - 1 \right| + 2\eta\sigma_* P_{\theta_*,i} \left| Z_i \right| \\ &\leq \frac{\sigma_*^2}{2} \left[\underbrace{P_{\theta_*,i}\, Z_i^2}_{=1} + 2 \left| \frac{r_*}{\sigma_*} \right| \underbrace{P_{\theta_*,i} \left| Z_i \right|}_{=\sqrt{2\pi^{-1}}} + 1 + 2\eta\sigma_* \underbrace{P_{\theta_*,i} \left| Z_i \right|}_{=\sqrt{2\pi^{-1}}} \right] \\ &\leq \sigma_*^2 \left[1 + \frac{\|\theta_*\|}{\sigma_*} \sqrt{2\pi^{-1}} + \eta\sigma_* \sqrt{2\pi^{-1}} \right] \triangleq c, \end{aligned}$$

qui permet d'obtenir le premier résultat, $P_{\theta_*,i}\, F_i(\theta_1, \delta) \leq \delta c$.

On note $d_i = c_i^2$ et $A_i = \left| Z_i^2 + 2 Z_i \frac{r_*}{\sigma_*} - 1 \right|$ et donc $P_{\theta_*,i}\, F_i(\theta_1, \delta)^2 \leq \delta^2\, P_{\theta_*,i}\, d_i$ avec

$$P_{\theta_*,i}\, d_i = \frac{\sigma_*^4}{4} P_{\theta_*,i} A_i^2 + 4\eta^2\sigma_*^2 P_{\theta_*,i} Z_i^2 + 2\eta\sigma_*^3 P_{\theta_*,i} |Z_i| \times A_i.$$

En appliquant l'inégalité $2ab \leq a^2 + b^2$ au terme $|Z_i| \times A_i$, on obtient

$$\begin{aligned} P_{\theta_*,i}\, d_i &\leq \frac{\sigma_*^4}{4} P_{\theta_*,i}\, A_i^2 + 4\eta^2\sigma_*^2 P_{\theta_*,i}\, Z_i^2 + 2\eta\sigma_*^3 \left(\frac{P_{\theta_*,i}\, Z_i^2}{2} + \frac{P_{\theta_*,i}\, A_i^2}{2} \right) \\ &= \frac{\sigma_*^4}{4} P_{\theta_*,i}\, A_i^2 + 4\eta^2\sigma_*^2 + 2\eta\sigma_*^3 \left(\frac{1}{2} + \frac{P_{\theta_*,i}\, A_i^2}{2} \right) \\ &= P_{\theta_*,i}\, A_i^2 \left(\frac{\sigma_*^4}{4} + \eta\sigma_*^3 \right) + 4\eta^2\sigma_*^2 + \eta\sigma_*^3. \end{aligned}$$

Avec quelques calculs, on trouve $P_{\theta_*,i}\, A_i^2 = 2(1 + 2\, r_*^2/\sigma_*^2) \leq 2(1 + 2\, \|\theta_*\|^2/\sigma_*^2)$ alors $\exists\, d > 0$ tel que $P_{\theta_*,i}\, d_i \leq d$, où d ne dépend que de η et θ_*. D'où, on en déduit le second résultat. □

Lemme 4.7. *Sous l'hypothèse 4, il existe $a > 0$ tel que pour tout $\theta \in \Theta$ et $i \geq 1$,*

$$P_{\theta_*,i}\, W_i(\theta, y_i)^2 \leq a.$$

Démonstration. On rappelle que $Z_i = (y_i - r_*(x_i))\,/\sigma_*$ et $W_i(\theta, y_i) = \frac{\sigma_*^2}{2\sigma^2}\,[Z_i^2 + 2 Z_i \Delta_i(r) - 1]$. D'où

$$W_i(\theta, y_i)^2 = \left(\frac{\sigma_*^2}{2\sigma^2} \right)^2 \left[Z_i^4 + 4 Z_i^3 \Delta_i(r) + 4 \Delta_i(r)^2 Z_i^2 - 4 Z_i \Delta_i(r) + 1 \right].$$

Sous $P_{\theta_*,i}$, $Z_i \sim \mathcal{N}(0,1)$, alors nous avons :

$$\begin{aligned} P_{\theta_*,i}\, W_i(\theta, y_i)^2 &= \left(\frac{\sigma_*^2}{2\sigma^2} \right)^2 \left[P_{\theta_*,i}\, Z_i^4 + 4\Delta_i(r) P_{\theta_*,i}\, Z_i^3 + 4\Delta_i(r)^2 P_{\theta_*,i}\, Z_i^2 - 4\Delta_i(r) P_{\theta_*,i}\, Z_i + 1 \right] \\ &= \frac{\sigma_*^4}{\sigma^4} \Delta_i(r)^2 \\ &\leq \sigma_*^2\, \|\theta\|^2\, \|\theta - \theta_*\|^2. \end{aligned}$$

Comme Θ est compact par l'hypothèse 4, il existe un δ_* qui majore $\|\theta - \theta_*\|$. De plus, il existe $\eta > 0$ tel que $\sup_{\theta\in\Theta} \|\theta\| \leq \eta$. Le résultat est alors démontré avec

$$a = \sigma_*^2 \eta^2 \delta_*^2.$$

□

Lemme 4.8. *Soit $\theta_0 = (\mu_0, \beta_0, \sigma_0^2) \in \Theta$, où β_0 est donné par l'hypothèse 3 et quelque soit $\mu_0 \in \mathbb{R}$ et $\sigma_0^2 \in \mathbb{R}_+^*$. Pour tout voisinage $U \subset \Theta$ de θ, $\Pi(U) > 0$.*

Démonstration. Notons que la boule $\{\theta \in \Theta : \|\theta - \theta_0\| \leq \varepsilon\}$ contient le produit cartésien

$$U_\mu \times U_\beta \times U_\tau = \{|\mu - \mu_0| \leq \varepsilon/3\} \times \{\|\beta - \beta_0\|_{L^2} \leq \varepsilon/3\} \times \{|\tau - \tau_0| \leq \varepsilon/3\}.$$

On rappelle que la distribution *a priori* est donnée par (2.9) dans le chapitre 2. On constate que les distributions *a priori* de μ et τ chargent U_μ et U_τ. La distribution *a priori* sur $\mathcal{E}_K$ se décompose sur $3K$ paramètres : $b_k \in \mathbb{R}$, $m_k \in [0, 1]$ et $\ell_k \in \mathbb{R}_+$, pour $k = 1, \ldots, K$. Soient b_{0k}, m_{0k} et ℓ_{0k}, les coefficients de la fonction β_0. Une faible variation autour de ces coefficients n'induit qu'une faible variation sur la norme L^2 de β_0, voir l'annexe 4.7.1 pour les détails techniques. Comme la distribution *a priori* (2.9) charge les voisinages de chacun des b_k, des m_k et des ℓ_k, elle charge $\{\beta \in \mathcal{E}_K : \|\beta_0 - \beta\| \leq \varepsilon/3\}$. □

4.6 Discussions

Avec le théorème 4.3, nous avons montré la consistance de la distribution *a posteriori* du modèle Bliss en $\theta_0 \in \Theta$. Nous avons déduit une caractérisation de θ_0 en fonction du *design* et du *vrai* paramètre θ_*. En particulier, nous avons établit que θ_0 est le paramètre le plus proche de θ_* au sens de la divergence de Kullback-Leibler. En ce qui concerne la fonction coefficient β_0, elle est la plus proche de β_* au sens de F. En montrant que F est convexe, nous montrons que si nous ne nous restreignons pas à l'ensemble des fonctions en escalier, une solution existe. De plus, nous avons remarqué que si β_* est une fonction en escalier, avec K escaliers ou moins, alors la distribution *a posteriori* de θ se concentre autour du vrai paramètre θ_*. Dans un cadre plus général, lorsque la *vraie* fonction n'est pas en escalier, nous quantifions avec la proposition 4.1 l'erreur alors commise sur les paramètres μ_* et σ_*^2 en fonction de l'erreur commise sur β_*.

Parmi les perspectives de travail possibles, la première serait d'étudier F sur l'ensemble $\mathcal{E}_K$ afin d'établir si l'hypothèse 3 est vérifiée dans notre cas ou si on peut la réduire à une hypothèse plus intuitive ou moins forte. Étant donné la topologie de $\mathcal{E}_K$ (voir le chapitre 2), il semble falloir se restreindre à l'étude d'un sous-ensemble de $\mathcal{E}_K$.
Pour établir simplement le résultat, nous avons supposé que l'espace paramétrique était compact. Pour relâcher cette hypothèse, Wald (1949) suggère des pistes (voir Ghosh and Ramamoorthi, 2003 pour une reformulation). Une perspective serait d'étudier ces suggestions afin de montrer si elles sont vérifiées dans notre contexte ou si des hypothèses plus faibles sont possibles.
Une autre perspective serait de considérer que le *design* est aléatoire. Avec un *design* aléatoire, on travaille avec les données $\{y_i, x_i\}$ pour $i = 1, \ldots, n$ qu'on pourra considérer comme indépendantes et identiquement distribuées selon une *vraie* loi P_{θ_*}. Nous ne

devrions alors plus avoir besoin du théorème de Kolmogorov en nous contentant d'une loi forte des grands nombres avec des hypothèses plus faibles à vérifier. En ce sens, nous avons l'intuition que le *design* aléatoire permettrait d'avoir un cadre plus simple pour montrer la consistance. Une hypothèse raisonnable sur le *design* dans ce contexte serait : "Pour tout $t \in [0,1]$, les données $x_1(t), \dots, x_n(t)$ sont indépendantes et identiquement distribuées et telles que $\mathbb{E}\, x_i(t)^2 < \infty$". Cette hypothèse implique que $\mathbb{E}\,|x_i(t)| < \infty$ et $\mathbb{E}\,|x_i(t)x_i(t')| < \infty$ si bien que par la loi forte des grands nombres, l'hypothèse 2 est vérifiée. Ce dernier point implique de plus que $c - e^{\otimes}$ est la fonction de covariance des x_i et est donc semi-définie positive. Ainsi, on trouve aussi que F est convexe dans ce contexte. Ces premiers points nous font suspecter que le théorème 4.3 devrait s'adapter au cas d'un *design* aléatoire.

Dans la même optique, une autre perspective de travail serait d'appliquer les travaux de Kleijn and van der Vaart (2006) qui ont étendu le théorème de Schwartz aux modèles non-paramétriques mal spécifiés. La consistance est donnée avec une hypothèse sur la distribution *a priori* qui semble être vérifiée dans notre cas, et une contrainte sur la métrique entropique du modèle. Cette entropie est une mesure de la complexité du modèle et permet généralement d'obtenir des vitesses de convergence de la distribution *a posteriori*.

Une autre perspective serait de montrer la consistance de l'estimateur du support de la fonction coefficient, qui est un des points importants de notre approche (voir les chapitres 1 et 2). Comme base de réflexion, on peut s'inspirer des travaux de Castillo et al. (2015) qui étudient la consistance du support de β dans un cadre de régression en dimension finie. Un dernière perspective, proche de la précédente, serait de montrer la consistance du signe de la fonction coefficient (voir Kang et al., 2016).

4.7 Annexes

4.7.1 Loi *a priori* et voisinage de β_0

Soit $\beta_0 \in \mathcal{E}_K$ caractérisé par $3K$ coefficients : b_{0k}, m_{0k} et ℓ_{0k} pour $k \leq K$. On note par $V_{\beta_0}(\delta)$ les fonctions β de $\mathcal{E}_K$ telles que pour tout $k \leq K$:

$$|b_{0k} - b_k| < \delta, \quad |m_{0k} - m_k| < \delta \text{ et } \quad |\ell_{0k} - \ell_k| < \delta$$

Si les intervalles d'une fonction $\beta \in \mathcal{E}_K$ admettent $2K$ bornes distinctes entre elles et qu'elles sont différentes de 0 et 1, ces bornes séparent $[0,1]$ en $2K+1$ zones. Si ce n'est pas le cas, le nombre de zones est inférieur à $2K+1$. Ces zones définissent une partition de $[0,1]$ et on note les bornes de ces zones par

$$0 = a_1 < a_2 < \cdots < a_{L-1} < a_L = 1,$$

avec $L \leq 2K+1$. On peut donc réécrire toute fonction β_0 et $\beta \in \mathcal{E}_K$ comme

$$\beta_0(t) = \sum_{k=1}^{L_0} c_{0k} \mathbf{1}_{[a_{0k}, a_{0k+1}[}(t) \text{ et } \quad \beta(t) = \sum_{k=1}^{L} c_k \mathbf{1}_{[a_k, a_{k+1}[}(t). \tag{4.10}$$

Par la suite, supposons que $L_0 = 2K+1$. On note $\Delta = \min_k |a_{0k+1} - a_{0k}| / 8$ et soit $\delta < \Delta$. On montre avec le lemme 4.9 que si $L_0 = 2K+1$, alors $L = L_0$ pour tout $\beta \in V_{\beta_0}(\delta)$. D'où pour tout $\beta \in V_{\beta_0}(\delta)$,

$$\begin{aligned}
\|\beta - \beta_0\|_{L^2}^2 &= \int_0^1 \left(\sum_{k=1}^{L} c_k \mathbf{1}_{[a_k, a_{k+1}[}(t) - \sum_{k=1}^{L} c_{0k} \mathbf{1}_{[a_{0k}, a_{0k+1}[}(t) \right)^2 \mathrm{d}t \\
&= \sum_{k=1}^{L} \int_0^1 \left(c_k \mathbf{1}_{[a_k, a_{k+1}[}(t) - c_{0k} \mathbf{1}_{[a_{0k}, a_{0k+1}[}(t) \right)^2 \mathrm{d}t \\
&\quad + \int_0^1 \sum_{k=1}^{L} \sum_{j \neq k} \left(c_k \mathbf{1}_{[a_k, a_{k+1}[}(t) - c_{0k} \mathbf{1}_{[a_{0k}, a_{0k+1}[}(t) \right) \times \left(c_j \mathbf{1}_{[a_j, a_{j+1}[}(t) - c_{0j} \mathbf{1}_{[a_{0j}, a_{0j+1}[}(t) \right) \mathrm{d}t \\
&= \sum_{k=1}^{L} A_k + \int_0^1 B(t) \, \mathrm{d}t.
\end{aligned}$$

On note λ la mesure de Lebesgue sur $[0,1]$. Pour majorer les A_k, on décompose le carré et on obtient

$$\begin{aligned}
A_k = &(c_k - c_{0k})^2 \, \lambda\Big([a_k, a_{k+1}[\cap [a_{0k}, a_{0k+1}[\Big) \\
&+ c_k^2 \, \lambda\Big([a_k, a_{k+1}[\setminus [a_{0k}, a_{0k+1}[\Big) \\
&+ c_{0k}^2 \, \lambda\Big([a_{0k}, a_{0k+1}[\setminus [a_k, a_{k+1}[\Big)
\end{aligned}$$

Le premier terme est majoré par $(c_k - c_{0k})^2$. On majore le deuxième terme en encadrant la valeur de c_k. En effet, comme c_k (resp. c_{0k}) est la somme (d'une partie) des b_k (resp. b_{0k}), alors $c_k \in [c_{0k} \pm \delta L]$. De plus, par le lemme 4.9 on a que $a_k \in [a_{0k} \pm 2\delta]$, d'où $\lambda\Big([a_k, a_{k+1}[\setminus [a_{0k}, a_{0k+1}[\Big) \leq |a_{0k} - a_k| + |a_{0k+1} - a_{k+1}| \leq 4\delta$. Pour majorer le dernier terme, on procède de la même manière que le point précédent. Il en résulte qu'il existe une constante $C_1 > 0$ telle que $\sum_{k=1}^{L} A_k \leq \delta C_1$.

Pour majorer $\int_0^1 B(t)\,\mathrm{d}t$, on va traiter chacun des termes issus de la décomposition du double produit. Comme les intervalles $[a_k, a_{k+1}[$ et $[a_j, a_{j+1}[$ sont disjoints pour $k \neq j$, certains termes de la décomposition sont nuls, de même pour les intervalles relatifs à β_0. De plus, puisque $\delta < \Delta$ et $a_k \in [a_{0k} \pm 2\delta]$, les intersections $[a_k, a_{k+1}[\cap[a_{0j}, a_{0j+1}[$ sont potentiellement non vides seulement si $j \in \{k-1, k, k+1\}$. Comme $j = k$ est exclu du double produit du faite de la somme $\sum_{j\neq k}$, on obtient après ces simplifications :

$$B(t) = -\sum_{k=1}^{L} \sum_{j\in\{k-1,k+1\}} c_k c_{0j} \mathbf{1}_{[a_k,a_{k+1}[\cap[a_{0j},a_{0j+1}[}(t).$$

En utilisant les mêmes majorations pour c_k dans le cas des A_k et pour celles concernant la mesure de Lebesgue des intersections d'intervalles, on trouve un $C_2 > 0$ tel que

$$\int_0^1 B(t)\,\mathrm{d}t \leq \delta C_2.$$

D'où pour tout $\varepsilon > 0$, il existe δ suffisamment petit tel que pour tout $\beta \in V_{\beta_0}(\delta)$, $\|\beta_0 - \beta\|_{L^2} < \varepsilon$.

Lemme 4.9. *Soient $\delta < \Delta$ et $\beta \in V_{\beta_0}(\delta)$, si $L_0 = 2K + 1$ alors $L = L_0$. De plus, pour tout $k = 1, \ldots, L$, $|a_k - a_{0k}| \leq 2\delta$.*

Démonstration. Comme $\beta \in V_{\beta_0}(\delta)$, on a pour tout $j = 1, \ldots, K$:

$$|(m_j - \ell_j) - (m_{0j} - \ell_{0j})| < 2\delta \quad \text{et} \qquad |(m_j + \ell_j) - (m_{0j} + \ell_{0j})| < 2\delta. \tag{4.11}$$

Comme $\delta < \Delta$, alors $|a_{0k} - a_{0k+1}| > 8\delta$ pour tout $j = 1, \ldots, K$. Comme chaque a_k est une borne $m_j - \ell_j$ ou $m_j + \ell_j$, la condition (4.11) implique que $a_{0k-1} < a_k < a_{0k+1}$. Il y a donc le même nombre de bornes distinctes si bien que les nombres de zones sont les mêmes : $L = L_0$. Ce dernier point implique avec (4.11) que $a_k \in [a_{0k} \pm 2\delta]$. □

On a ici supposé que $L_0 = 2K+1$ pour ne pas surcharger la démonstration du résultat de détails techniques mais le résultat reste vrai si on ne fait pas cette hypothèse. Celle-ci revient à négliger les cas où des bornes de deux intervalles $\mathcal{I}_{0k}$ et $\mathcal{I}_{0j}$ sont égales à un $c \in]0, 1[$, ou le cas où une borne d'un intervalle $\mathcal{I}_{0k}$ vaut 0 ou 1. Le premier cas est négligeable par rapport à la loi *a priori* sur les m_k et les ℓ_k. Le second cas arrive lorsque pour au moins un k, $m_{0k} + \ell_{0k} > 1$ ou $m_{0k} - \ell_{0k} < 0$. Or ce second cas n'est pas négligeable par rapport à la loi *a priori*. Il suffit alors de choisir δ suffisamment petit afin que pour tout les k tels que $m_{0k} + \ell_{0k} > 1$, on ait $m_k + \ell_k > 1$ (et symétriquement pour la borne 0). En particulier, le résultat du lemme 4.9 reste vrai avec $L_0 \neq 2K + 1$: $L = L_0$ et $a_k \in [a_{0k} \pm 2\delta]$.

V

Perspectives

Dans le cadre de cette thèse, nous avons répondu à des attentes pratiques, notamment dans le domaine de l'agronomie. Non seulement nous avons proposé une méthode dans le cadre d'un modèle de régression linéaire fonctionnel fournissant des estimateurs simples avec une notion de crédibilité mais nous avons étendu la méthode pour prendre en compte efficacement l'avis d'experts. De plus, nous avons validé ces méthodes avec un résultat de consistance de la distribution *a posteriori*. De chacun de ces trois chapitres, de nouvelles perspectives de travail émergent.

Concernant les développements méthodologiques de la méthode Bliss dans le chapitre 2, plusieurs perspectives de travail émergent, notamment pour répondre à des questions agronomiques. Une d'entre elles serait de considérer un modèle avec une variable qualitative et une fonction coefficient pour chacun des niveaux de cette variable supplémentaire. Cette extension permettrait d'étudier la réaction d'une plante face au climat en fonction de la variété de la plante ou de son génotype. L'objectif serait de déterminer quelle variété utiliser suivant le climat de la région. Cette extension ne présente pas de difficulté majeure d'un point de vue méthodologique, mais demande des efforts d'implémentation.
De plus, dans certains contextes agronomiques, on peut vouloir prendre en compte cette variable qualitative comme un effet aléatoire. Par exemple, si on modélise le rendement du maïs, mesuré sur plusieurs parcelles similaires, on peut considérer les variations entre parcelles comme un effet aléatoire. Une extension importante est donc le modèle Bliss mixte.
Un autre extension importante pour les applications en agronomie serait de modéliser une corrélation entre les y_i. Par exemple, lorsque y_i est la teneur en sucre des raisins d'un pied de vigne, il faut prendre en compte une corrélation spatiale entre les y_i, surtout pour les pieds de vigne proches sur la parcelle. Dans un autre cas, si les y_i correspondent au rendement d'une parcelle sur différentes années, une corrélation temporelle peut être envisagée. Cette corrélation permettrait, par exemple, de modéliser qu'une année avec un haut rendement peut induire un faible rendement l'année suivante.
D'autres types de données pourraient être modélisés si des variantes du modèle Bliss

étaient développées. Par exemple, un modèle Bliss généralisé permettrait notamment d'étudier les cas où y_i n'est pas un rendement, mais par exemple à une présence ou une absence. On peut aussi considérer un autre type de données importantes dans le cadre de l'application du modèle de régression linéaire fonctionnel en agronomie : les données cycliques. Par exemple, si $x_i(\cdot)$ est la température, qu'on mesure toutes les heures pendant plusieurs jours, la redondance dans les données doit être prise en compte dans le modèle.
Une autre perspective importante pour des applications en agronomie serait de développer un modèle Bliss qui prenne en compte des données $x_i(\cdot)$ qui n'ont pas forcément le même domaine. En effet, pour que les données fonctionnelles soient comparables, il est souvent nécessaire de considérer une dilatation temporelle pour chacun des $x_i(\cdot)$.
Une dernière attente importante des agronomes dans le cas du modèle Bliss est de pouvoir déterminer les contributions respectives de chacune des covariables fonctionnelles dans le modèle.

D'un point de vue modélisation, une autre perspective importante serait d'inclure un terme d'interaction entre deux covariables fonctionnelles. Lorsque les covariables sont fonctionnelles, un terme d'interaction entre $\beta_i(\cdot)$ et $\beta_j(\cdot)$ (dit terme quadratique) est introduit par un terme bidimensionnel $\beta_{ij}(\cdot,\cdot)$. Des modèles de régression fonctionnels incluant un tel terme ont été étudié par Yao and Müller (2010); Yang et al. (2013); McLean et al. (2014); Fuchs et al. (2015); Scheipl et al. (2015); Usset et al. (2016); Greven and Scheipl (2017) avec en particulier des méthodes permettant de tester la significativité d'un terme quadratique (Horváth and Kokoszka, 2012; Zhang et al., 2014). Cette perspective se rapproche d'une autre extension possible : étendre le modèle Bliss aux covariables fonctionnelles de dimension supérieure, comme les images 2D (Huang et al., 2013) ou les images 3D (Wang et al., 2014). Dans cette optique, on peut souhaiter estimer le terme quadratique par une fonction parcimonieuse et constante par morceaux. Se pose alors la question de la forme des fonctions constantes par morceaux en dimension supérieure. En particulier pour la dimension 2, l'extension naïve serait de définir un intervalle en deux dimensions comme le produit cartésien de deux intervalles de dimension 1. On aurait alors deux milieux et deux demi-longueurs pour caractériser un tel "intervalle". Cependant, est-ce que cette forme serait adéquate pour estimer un terme quadratique ?
Une autre extension possible serait de considérer un *design* aléatoire pour la covariable fonctionnelle $x(\cdot)$. Un des atouts de cette approche est que nous pourrions alors tenter de répondre à une question importante des agronomes. En effet, au-delà de la prédiction d'une production y à partir d'un scénario climatique $x(\cdot)$, ou de l'interprétation de la fonction coefficient, l'objectif pour un agronome serait de connaître les scénarios climatiques pour lesquels la production y serait plus grande qu'un seuil τ. Autrement dit, on pourrait s'intéresser à la distribution $\pi\left(x(\cdot) \mid y > \tau\right)$.
Enfin, un travail important serait le développement d'un package R comprenant la méthode Bliss et ses extensions. L'objectif serait qu'un agronome ayant de bonnes connaissances en statistiques puisse lui-même appliquer la méthode Bliss.

Pour ce qui est du chapitre 3, deux principales perspectives se dégagent. Premièrement, dans certains cas la calibration des poids proposée en section 3.2.3 peut ne pas être satisfaisante. En effet, imaginons le cas où l'ensemble des experts ne fournissent que très peu de données, alors le terme de renormalisation $\frac{n}{n_e \times E}$ gonflerait artificiellement le poids des experts. Une deuxième critique est que la détermination de la dépendance entre les experts est laissé à l'attention du statisticien alors que ce terme peut influencer l'inférence.

Une piste serait d'établir une procédure pour estimer cette dépendance lors d'une réunion avec les experts.
La deuxième perspective concerne l'approche par pénalisation. Si ce terme de pénalisation peut être attrayant pour prendre en compte l'avis des experts puisqu'on trouve une interprétation du paramètre d'intensité τ, déterminer τ s'avère complexe. Pour le calibrer, l'objectif serait de trouver un équivalent échantillon, comme l'hyperparamètre g du *g-prior* de Zellner. Ainsi, on pourrait choisir τ comme un "demi-échantillon", si bien que l'avis des experts conterait au même titre que la moitié des données observées dans l'inférence.

Plusieurs perspectives sont envisageables concernant le résultat du chapitre 4. La premier concerne l'hypothèse de compacité de l'espace paramétrique. Cette hypothèse n'est pas vérifiée dans notre contexte, mais il existe des solutions pour la relâcher, voir par exemple Wald (1949) et Ghosh and Ramamoorthi (2003).
Une autre perspective serait d'établir un résultat général pour la consistance du modèle de régression linéaire fonctionnel avec des hypothèses différentes de celles proposées par Lian et al. (2016).
Enfin, une dernière perspective possible serait de s'inspirer des travaux de Amewou-Atisso et al. (2003); Kleijn and van der Vaart (2006); Choi (2009) ou Xiang and Walker (2013) pour démontrer une version du théorème de Schwartz dans notre contexte.

VI

Annexe

6.1 Exemple de mise en œuvre de la méthode Bliss

Dans cette section, nous présentons une exemple d'application de la méthode Bliss en utilisant un code disponible en ligne à l'adresse :
http://www.math.univ-montp2.fr/~grollemund/Bliss/
Le code disponible comprend l'implémentation de tout les développements méthodologiques de cette thèse. L'ensemble des fonctions est implémenté avec le langage *R*, avec en particulier l'utilisation du package *Rcpp* pour les fonctions demandant beaucoup de calculs.

Dans ce qui suit, nous donnons un exemple qui détaille pas à pas les informations nécessaires pour mettre en œuvre la méthode Bliss, présentée dans le chapitre 2. En particulier nous expliquons comment :

- simuler des données pour illustrer le modèle Bliss,
- obtenir un échantillon de la distribution *a posteriori* avec un échantillonneur de Gibbs,
- représenter la distribution *a posteriori* de la fonction coefficient et celle de son support,
- calculer les différents estimateurs bayésiens et tracer les résultats dans un graphique.

Nous présentons ici un exemple avec une seule covariable fonctionnelle, mais l'implémentation prend en charge le cas avec plusieurs covariables fonctionnelles. Des fonctionnalités, comme le calcul du critère BIC pour le choix de K ou des diagnostiques de convergence de l'échantillonneur de Gibbs, sont données en ligne mais ne sont pas détaillées ici. L'implémentation des méthodes présentées dans le chapitre 3 est également disponible en ligne. Leurs mises en œuvre étant en grande partie identique à celle de la méthode Bliss, nous n'en présentons pas d'exemple d'application.

Simulation de données Pour obtenir des données, nous utilisons un schéma de simulation détaillé en section 2.3.1. Plusieurs caractéristiques sont à spécifier pour simuler des données, comme `n` (le nombre d'observations), `p` (le nombre d'instants de mesure des courbes $x_i(\cdot)$), `beta_types` (la forme de la fonction coefficient), ainsi que `b_inf` et `b_sup` (pour définir le domaine des courbes $x_i(\cdot)$). A partir de ces paramètres, on utilise dans le code suivant, la fonction `sim_multiple` pour simuler des courbes $x_i(\cdot)$ et des valeurs réelles y_i, à partir du modèle de régression linéaire fonctionnel. On donne une représentation des données ainsi obtenues dans la figure 6.1.

```
param_sim ← list(n=25,p=15,beta_types="smooth",b_inf=0,b_sup=1)
data ← sim_multiple(param_sim)
# Simulation of the data.
#   Simulate the functions x_qi(t).
#   Choose a coefficient function.
#   Compute the outcomes y_i.
```

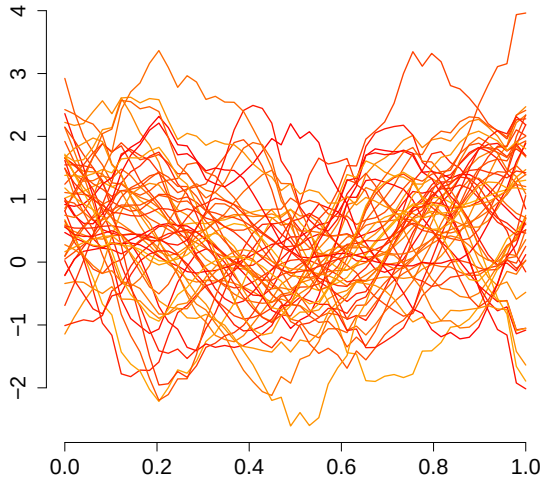


Figure 6.1: Représentation des courbes $x_1(\cdot), \ldots, x_n(\cdot)$ issues du jeu de données simulées.

Echantillonner la distribution *a posteriori* Pour obtenir un échantillon *a posteriori*, nous utilisons l'algorithme de Gibbs dont des détails sont donnés en section 2.6.2. On utilise la fonction principale `Bliss_multiple` qui appelle des sous-fonctions permettant entres autres d'échantillonner la distribution *a posteriori*, puis de calculer la distribution *a posteriori* de la fonction coefficient, d'exécuter un algorithme d'optimisation pour calculer une estimation constante par morceaux, de calculer une estimation du support et de calculer les densités de l'échantillon *a posteriori* (utile pour calculer le critère BIC). Cette fonction principale nécessite une liste `param` contenant

`iter`, le nombre d'itérations de l'algorithme de Gibbs,

`burnin`, le temps de chauffe,

`K`, l'hyperparamètre K du modèle Bliss,

`grids`, les instants de mesure des courbes $x_i(\cdot)$,

`prior_beta`, un argument spécifiant la distribution *a priori* de β (seule l'option `"Ridge_Zellner"` est considérée dans ce manuscrit) et

`phi_l`, un argument spécifiant la distribution *a priori* de ℓ (seule l'option `"Gamma"` est considérée dans ce manuscrit).

```
param ← list(iter=5e4,burnin=1e3, K=3,grids=data$grids,
                  prior_beta="Ridge_Zellner",phi_l=list("Gamma"))
res_Bliss ← Bliss_multiple(data,param)
# Gibbs Sampler:
#   Initialization.
#   Determine the starting point.
#   Start the Gibbs loop.
#   Return the result.
# Compute the functions beta_i.
# Compute the estimation of the posterior density.
#   Thin the sample.
# Perform the 'kde2d' function.
# Simulated Annealing:
#   Initialization.
#   Determine the starting point.
#   Start the loop.
#   Return the result.
# Estimation of the support.
# Compute the (log) densities.
```

Représentation de la distribution *a posteriori* Nous donnons ici le code pour obtenir des représentations de la distribution *a posteriori*. En premier lieu, nous donnons le code pour obtenir les probabilités *a posteriori* $\alpha(t|\mathcal{D})$, relative au support (voir la section 2.2.4). Ensuite, la fonction `image_Bliss` est utilisé pour représenter la distribution *a posteriori* de la fonction coefficient (voir la section section 2.2.7 pour plus de détails). On donne ces résultats dans la figure 6.2.

```
#### Posterior probabilities alpha(t)
plot(data$grids[[1]],res_Bliss$alpha_t[[1]],type="l",ylim=c(0,1),xlab=""
   ,ylab="", axes=F)
axis(1,cex.axis=1.5)
axis(2,cex.axis=1.5)

#### Posterior distribution of the coefficient function
image_Bliss(res_Bliss$posterior_density_estimate[[1]],param)
```

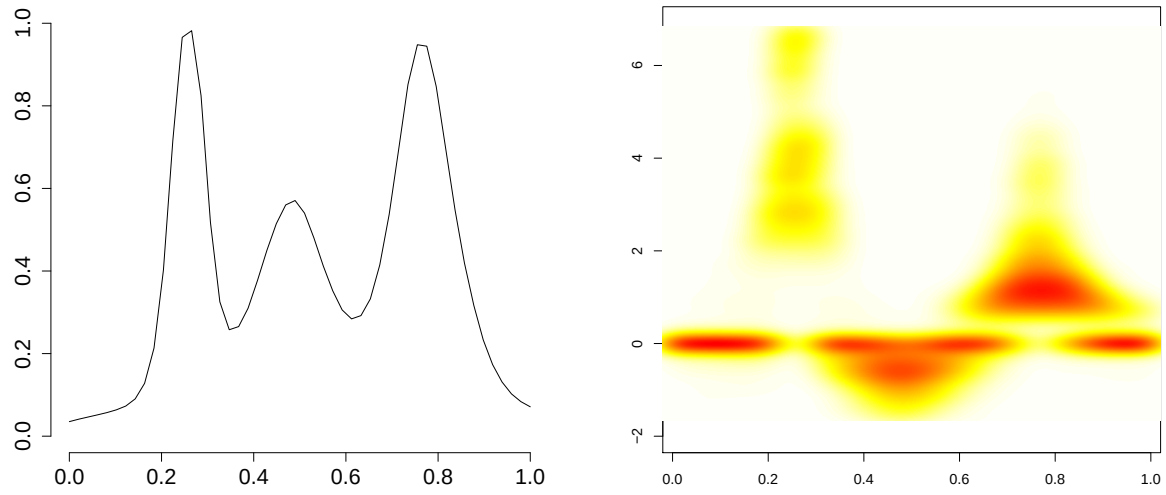


Figure 6.2: Représentation des distributions *a posteriori* de la fonction coefficient et de son support. *Le premier graphique est une représentation des probabilités a posteriori $\alpha(t|\mathcal{D})$, voir section 2.2.4 pour une description détaillée. Le second graphique est une représentation de la distribution a posteriori de la fonction coefficient. La couleur rouge (resp. blanche) est utilisée pour représenter une forte (resp. faible) valeur de la densité a posteriori.*

Calculer et représenter des estimateurs Nous donnons les lignes de code nécessaires pour obtenir l'estimateur du support, introduit en section 2.2.4, et les deux estimateurs de la fonction coefficient, introduit en section 2.2.5. Dans cet exemple, on estime le support avec trois intervalles, dont le premier commence en `data$grids[[1]][12]` et finit en `data$grids[[1]][16]`. Les estimations de la fonction coefficient sont ici données de manière vectorielle, telle que chaque élément est la valeur de l'estimation en un instant de la grille `data$grids[[1]]`.

```
#### Support estimator
res_Bliss$support_estimate[[1]]
#   begin end
# 1 12    16
# 2 23    26
# 3 35    42

#### Smooth estimator of the coefficient function
round(t(res_Bliss$res.Simulated_Annealing[[1]]$posterior_expe),3)
# [,1]   [,2]   [,3]   [,4]   [,5]   [,6]  [,7]  [,8]  [,9] [,10]  [,11]
# 0     -0.002 -0.004 -0.006 -0.004  0     0.011 0.037 0.119 0.366 1.092
# [,12] [,13] [,14] [,15] [,16] [,17] [,18]  [,19]  [,20]  [,21]  [,22]
# 2.832 4.618 4.759 3.67  1.588 0.509 0.089 -0.088 -0.184 -0.263 -0.346
# [,23]  [,24]  [,25]  [,26]  [,27]  [,28]  [,29]  [,30] [,31] [,32]
#-0.426 -0.492 -0.511 -0.47  -0.384 -0.286 -0.197 -0.11 -0.03  0.059
# [,33] [,34] [,35] [,36] [,37] [,38] [,39] [,40] [,41] [,42] [,43]
# 0.173  0.34  0.6  0.989 1.473 1.814 1.811 1.486 1.036 0.685 0.445
# [,44] [,45] [,46] [,47] [,48] [,49] [,50]
# 0.287 0.182 0.115 0.076 0.051 0.038 0.033

#### Bliss estimator : stepwise estimator of the coefficient function
```

```
round(t(res_Bliss$Bliss_estimate[[1]]),3)
# [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
# 0    0    0    0    0    0    0    0    0     0    0     3.219 3.219
# [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24]
# 3.219 3.219 3.219 0     0     0     0     0     0     0     0
# [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35]
# 0     0     0     0     0     0     0     0     0     0     0
# [,36] [,37] [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46]
# 1.887 1.887 1.887 1.887 1.887 0     0     0     0     0     0
# [,47] [,48] [,49] [,50]
# 0     0     0     0
```

Avec les lignes suivantes, on donne le code pour avoir une représentation de ces estimateurs avec la distribution *a posteriori* (voir la figure 6.3 pour les résultats graphiques).

```
#### Plot the support estimate
plot(data$grids[[1]],res_Bliss$alpha_t[[1]],type="l",ylim=c(0,1),xlab=""
   ,ylab="")
for(k in 1:nrow(res_Bliss$support_estimate[[1]])){
  segments(data$grids[[1]][res_Bliss$support_estimate[[1]][k,1]],0.5,
           data$grids[[1]][res_Bliss$support_estimate[[1]][k,2]],0.5,lwd
   =2,col=2)
  points(data$grids[[1]][res_Bliss$support_estimate[[1]][k,1]],0.5,pch="
   |",lwd=2,col=2)
  points(data$grids[[1]][res_Bliss$support_estimate[[1]][k,2]],0.5,pch="
   |",lwd=2,col=2)
}
abline(h=0.5,col=2,lty=2)

#### Plot the coefficient function estimate
image_Bliss(res_Bliss$posterior_density_estimate[[1]],param)
lines.step_function(res_Bliss$param$grids2 [[1]],
                    res_Bliss$Bliss_estimate[[1]],lwd=2,bound=F)
lines(res_Bliss$param$grids[[1]],
      res_Bliss$res.Simulated_Annealing[[1]]$posterior_expe,
      lty=2)
lines(data$grids[[1]],data$beta[[1]],col=3)
```

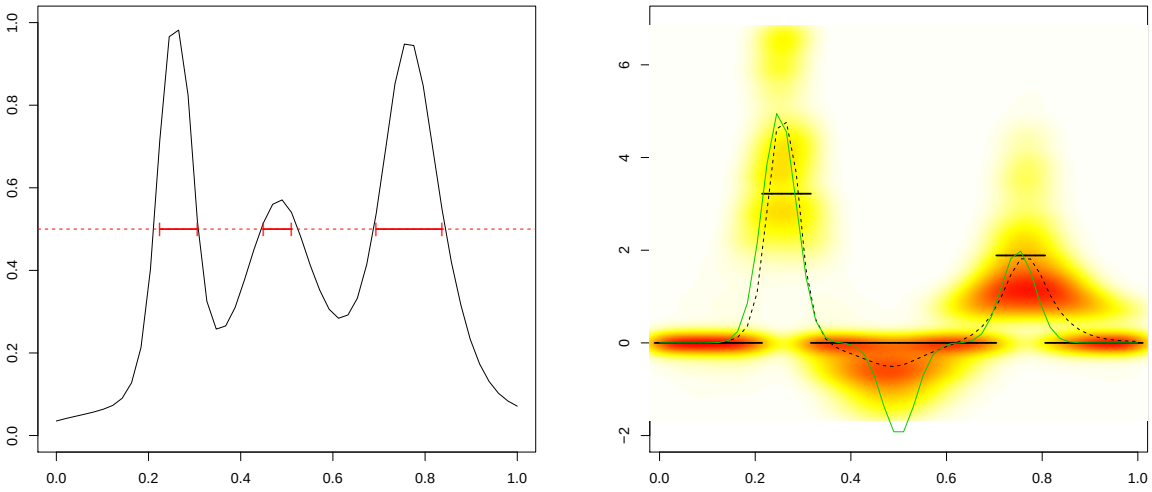


Figure 6.3: Représentation de l'estimateur du support et des estimateurs de la fonction coefficient. *Pour le premier graphique, l'estimateur du support est donné par les segments rouges. Pour le second graphique, l'estimation constante par morceaux de la fonction coefficient est donnée par la courbe noire en trait plein. L'estimateur lisse est donné par la courbe en pointillé. La courbe verte correspond à la vraie fonction coefficient utilisée pour générer les données.*

Bibliographie

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., and Rousseau, J. (2012). Combining Expert Opinions in Prior Elicitation. *Bayesian Analysis*, 7. 4, 56

Alpert, M. and Raiffa, H. (1982). *A progress report on the training of probability assessors*, pages 294–305. Cambridge University Press, Cambridge. 59

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., Ramamoorthi, R., et al. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9(2):291–312. 9, 94, 111

Anand, P. (1985). A critique of the normative and descriptive foundations of subjective probability - with reference to agriculture. *Journal of Economic Psychology*, 6(4):399 – 416. 54

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79. 13

Baragatti, M. and Pommeret, D. (2012). A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics and Data Analysis*, 56(6):1920–1934. 17

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005). Hierarchical models for assessing variability among functions. *Biometrika*, 92(2):419–434. 12

Bélisle, C. (1992). Convergence Theorems for a Class of Simulated Annealing Algorithms on $\mathbb{R}^d$. *Journal of Applied Probability*, 29(4):885–895. 46

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. Springer-Verlag. 59, 60

Berk, R. H. et al. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58. 94

Brown, P. J., Fearn, T., and Vannucci, M. (2001). Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem. *Journal of the American Statistical Association*, 96(454):398–408. 4, 12

Büntgen, U., Egli, S., Camarero, J., Fischer, E., Stobbe, U., Kauserud, H., Tegel, W., Sproll, L., and Stenseth, N. (2012). Drought-induced decline in Mediterranean truffle harvest. *Nature Climate Change*, 2:827–829. 38

Büntgen, U., Tegel, W., Egli, S., Stobbe, U., Sproll, L., and Stenseth, N. (2011). Truffles and climate change. *Frontiers in Ecology and the Environment*, 9(3):150–151. 35

Burgman, M., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L., and Twardy, C. (2011). Expert status and performance. *Statistical PLoS ONE*, 6. 55

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22. 12

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–591. 3, 12

Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*, volume 17. Chapman & Hall/CRC Boca Raton, FL. 56

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis*, 13(4):223–242. 63

Castillo, I. (2014). Bayésien non-paramétrique, convergence et forme limite de lois a posteriori. 94

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018. 106

Chen, M.-H. and Dey, D. (2003). Variable selection for multivariate logistic regression models. *Journal of Statistical Planning and Inference*, 111. 54

Chen, M.-H., Ibrahim, J. G., et al. (2006). The relationship between the power prior and hierarchical models. *Bayesian Analysis*, 1(3):551–574. 60

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media. 69

Choi, T. (2009). Asymptotic properties of posterior distributions in nonparametric regression with non-gaussian errors. *Annals of the Institute of Statistical Mathematics*, 61(4):835–859. 95, 111

Choi, T. and Schervish, M. J. (2004). Posterior consistency in nonparametric regression problems under gaussian process priors. 9, 94

Chu, H.-C. and Hwang, G.-J. (2008). A delphi-based approach to developing expert systems with the cooperation of multiple experts. *Expert Systems with Applications*, 34. 55

Cooke, R. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. 4, 55, 62, 63

Cooke, R. and Goosens, L. (2000). Procedures guide for structured expert judgment in accident consequence modelling. *Radiation Protection Dosimetry*, 90. 55

Crainiceanu, C. and Goldsmith, A. (2010). Bayesian functional data analysis using winbugs. *Journal of Statistical Software, Articles*, 32(11):1–33. 4, 12, 24

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software*, 14(14):1–24. 12

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1):35–72. 12, 23

Crowder, M. (1992). Bayesian priors based on a parameter transformation using the distribution function. *Annals of the Institute of Statistical Mathematics*, 44(3):405–416. 88

Dalkey, N. and Helmer, O. (1963). An experimental application of the delphi method to the use of experts. *Management Science*, 9. 55

De Groot, M. (1974). Reaching a consensus. *J. Amer. Statist. Assoc.*, 69:118–121. 63

De Santis, F. (2006). Power priors and their use in clinical trials. *The American Statistician*, 60(2):122–129. 60

Deheuvels, P. (1980). *L'intégrale*. Presse Universitaire de France. 23

Delbecq, A. and Van de Ven, A. (1971). A group process model for problem identification and program planning. *Journal of Applied Behavioral Science*, 7. 55

Demerson, J. and Demerson, M. (2014). *La truffe, la trufficulture, vues par les Demerson, Uzès (1989-2015)*. Les éditions de la Fenestrelle. 34, 38

Denham, R. and Mengersen, K. (2007). Geographically assisted elicitation of expert opinion for regression models. *Bayesian Analysis*, 2. 54

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458. 23

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27. 94

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102. 3

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media. 2, 12

Fleishman, E., Nally, R., Fay, J., and Murphy, D. (2001). Modeling and Predicting Species Occurrence Using Broad-Scale Environmental Variables: An Example with Butterflies of the Great Basin. *Conservation Biology*, 15. 55

French, S. (1985). Group consensus probability distributions: A critical survey. pages 183–201. 64

Fuchs, K., Scheipl, F., and Greven, S. (2015). Penalized scalar-on-functions regression with interaction term. *Computational Statistics & Data Analysis*, 81:38–51. 110

Garthwaite, P., Kadane, J., and O'Hagan, A. (2005a). Statistical methods for eliciting probability distributions. 54

Garthwaite, P. H. and Dickey, J. M. (1991). An elicitation method for multiple linear regression models. *Journal of Behavioral Decision Making*, 4(1):17–31. 56

Garthwaite, P. H. and Dickey, J. M. (1992). Elicitation of prior distributions for variable-selection problems in regression. *Ann. Statist.*, 20:1697–1719. 56

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005b). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701. 63

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Stanford University Department of Statistics. 69

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185. 18

Genest, C. and Zidek, J. (1986). Combining probability distributions. a critique and annoted bibliography. *Statistical Science*, 1. 55

Ghosal, S. (1997). A review of consistency and convergence of posterior distribution. In *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*. 94

Ghosal, S., Ghosh, J. K., Van Der Vaart, A. W., et al. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531. 94

Ghosal, S., Van Der Vaart, A., et al. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223. 9, 94

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer New York, New York, NY. 94, 105, 111

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. 63

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2010). Longitudinal penalized functional regression. 12

Goldsmith, J., Huang, L., and Crainiceanu, C. (2014). Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection. *J. Comput. Graph. Stat.*, 23(1):46–64. 13

Goldsmith, J., Wand, M. P., and Crainiceanu, C. (2011). Functional regression via variational Bayes. *Electronic journal of statistics*, 5:572–602. 4, 12

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114. 69

Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, 17:1–35. 110

Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58(1):121–128. 2

Hall, P. and Hooker, G. (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):637–653. 2

Healy, R., Smith, M., Bonito, G., Pfister, D., Ge, Z., Guevara, G., Williams, G., Stafford, K., Kumar, L., Lee, T., Hobart, C., Trappe, J., Vilgalys, R., and McLaughlin, D. (2013). High diversity and widespread occurence of mitotic spore mats in ectomycorrhizal Pezizales. *Molecular Ecology*, 22(6):1717–1732. 38

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. 66

Horváth, L. and Kokoszka, P. (2012). A test of significance in functional quadratic regression. *Inference for Functional Data with Applications*, pages 225–232. 110

Hu, F. and Zidek, J. (1995). Incorporating relevant sample information using the likelihood. Vancouver, BC. 7, 59, 61

Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *Canadian Journal of Statistics*, 30(3):347–371. 60

Huang, L., Goldsmith, J., Reiss, P. T., Reich, D. S., and Crainiceanu, C. M. (2013). Bayesian scalar-on-image regression with application to association between intracranial dti and cognitive outcomes. *NeuroImage*, 83:210–223. 110

Huber, G. P. (1974). Methods for quantifying subjective probabilities and multi-attribute utilities. *Decision Sciences*, 5(3):430–458. 56

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA. 94

Hunns, D. and Daniels, B. (1981). Paired comparisons and estimates of failure likelihood. *Design Studies*, 2. 55

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.*, 15(1):46–60. 7, 54, 59, 60, 61

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749. 59, 62

James, A., Low-Choy, S., and Mengersen, K. (2010). Elicitator: An expert elicitation tool for regression in ecology. *Environmental Modelling & Software*, 25. 56

James, G., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108. 3, 13, 24

Jenkinson, D. (2005). The elicitation of probabilities - A review of the statistical literature. *Technical Report*. 54

Kadane, J., Dickey, J., Winkler, R., Smith, W., and Peters, S. (1980). Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association*, 75(372):845–854. 55, 56

Kadane, J. and Wolfson, L. (1998). Experiences in elicitation. *The statistician*, 47. 4, 55, 56

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91. 59

Kang, J., Reich, B. J., and Staicu, A.-M. (2016). Scalar-on-image regression via the soft-thresholded gaussian process. *arXiv preprint arXiv:1604.03192*. 4, 13, 106

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, 220(4598):671–680. 23

Kleijn, B. (2016). On the frequentist validity of bayesian limits. *arXiv preprint arXiv:1611.08444*. 94

Kleijn, B. J. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, pages 837–877. 9, 94, 106, 111

Krinitzsky, E. L. (1993). Earthquake probability in engineering - part 1: The use and misuse of expert opinion. the third richard h. jahns distinguished lecture in engineering geology. *Engineering Geology*, 33(4):257 – 288. 54

Kuhnert, P., Martin, T., Mengersen, K., and Possingham, H. (2005). Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. *Environmetrics*. 56

Kynn, M. (2005). *Eliciting Expert Knowledge for Bayesian Logistic Regression in Species Habitat Modelling*. PhD thesis, Queensland University of Technology. 4, 54, 58

Kynn, M. (2008). The "heuristics and biaises" bias in expert elicitation. *J. R. Statist. Soc. A*, 171. 4, 55, 63

Le Tacon, F., Marçais, B., Courvoisier, M., Murat, C., Montpied, P., and Becker, M. (2014). Climatic variations explain annual fluctuations in French Périgord black truffle wholesale markets but do not explain the decrease in black truffle production over the last 48 years. *Mycorrhiza*, 24:S115–S125. 34, 35, 38, 80

Le Tacon, F., Rubini, A., Murat, C., Riccioni, C., Robin, C., Belfiori, B., Zeller, B., De La Varga, H., Akroume, E., Deveau, A., Martin, F., and Paolocci, F. (2016). Certainties and uncertainties about the life cycle of the Périgord black Truffle (Tuber melanosporum Vittad.). *Annals of Forest Science*, 73(1):105–117. 35, 38, 80

Li, F., Zhang, T., Wang, Q., Gonzalez, M., Maresh, E., and Coan, J. (2015). Spatial Bayesian Variable Selection and Grouping for High-Dimensional Scalar-on-Image Regression. *The Annals of Applied Statistics*, 23(2):687–713. 13

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327. 3

Lian, H., Choi, T., Meng, J., and Jo, S. (2016). Posterior convergence for bayesian functional linear regression. *Journal of Multivariate Analysis*, 150:27 – 41. 95, 111

Lindley, D. V. (1982). Scoring rules and the inevitability of probability. *International Statistical Review/Revue Internationale de Statistique*, pages 1–11. 63

Low Choy, S. (2012). *Priors: Silent or Active Partners of Bayesian Inference?*, pages 30–65. John Wiley & Sons, Ltd. 4

Low-Choy, S., James, A., and Mengersen, K. (2009). Expert elicitation and its interface with technology: a review with a view to designing Elicitator. 55

Marin, J.-M. and Robert, C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 62

Marin, J.-M. and Robert, C. (2010). *Importance sampling methods for Bayesian discrimination between embedded models*, chapter 14, pages 513–527. Springer-Verlag, New York. 18

Marin, J.-M. and Robert, C. P. (2009). Importance sampling methods for bayesian discrimination between embedded models. *arXiv preprint arXiv:0910.2325*. 69

Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93(442):740–750. 62

Martin, T., Kuhnert, P., Mengersen, P., and Possingham, H. (2005). The power of expert opinion in ecological models: a bayesian approach examining the impact of livestock grazing on birds. *Ecological Applications*, 15. 56

Marx, B. D. and Eilers, P. H. (1999). Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41(1):1–13. 3

Marx, B. D. and Eilers, P. H. (2005). Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22. 2

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269. 110

Meyer, M. and Booker, J. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. 4, 54

Montagna, S., Tokdar, S. T., Neelon, B., and Dunson, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, 68(4):1064–1073. 12

Morris, J. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359. 2, 12

Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240. 2

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, pages 774–805. 2

Murat, C., Rubini, A., Riccioni, C., De La Varga, H., Akroume, E., Belfiori, B., Guaragno, M., Le Tacon, F., Robin, C., Halkett, F., Martin, F., and Paolocci, F. (2013). Fine-scale spatial genetic structure of the black truffle (Tuber Melanosporum) investigated with neutral microsatellites and functional mting type genes. *The New Phytologist*, 199(1):176–187. 38

Murphy, A. H. and Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta psychologica*, 34:273–286. 63

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Published Examples of the Formal Elicitation of Expert Opinion*, pages 193–216. John Wiley & Sons, Ltd. 4, 54, 56, 59, 88

O'Leary, R., Low-Choy, S., Murray, J., Kynn, M., Denham, R., Martin, T., and Mengersen, K. (2009). Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby petrogale penicillata. *Environmetrics*, 20. 54

O'Leary, R., Mengersen, K., and Low-Choy, S. (2008). A mixture model approach to representing simple expert information as priors in logistic regression. *Technical report, School of Mathematical Sciences, Queensland University of Technology.* 55, 56

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518. 3

Ouchi, F. (2004). A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis. 4, 54, 55

Pan, W., Xie, B., and Shen, X. (2010). Incorporating Predictor Network in Penalized Regression with Application to Microarray Data. *Biometrics.* 54

Park, A. Y., Aston, J. A., and Ferraty, F. (2016). Stable and predictive functional domain selection with application to brain images. *arXiv preprint arXiv:1606.02186.* 3

Petrone, S., Guindani, M., and Gelfand, A. E. (2009). Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):755–782. 12

Phythian, J. E. and Williams, R. (1986). Direct cubic spline approximation to integrals with applications in nautical science. *International Journal for Numerical Methods in Engineering*, 23:305–315. 23

Picheny, V., Servien, R., and Villa-Vialaneix, N. (2016). Interpretable sparse sir for functional data. *arXiv preprint arXiv:1606.00614.* 3, 13

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis.* Springer-Verlag New York. 2

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis.* Springer-Verlag New York. 12, 23

Ray, S. and Mallick, B. (2006). Functional clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):305–332. 12

Reiss, P., Goldsmith, J., Shang, H., and Ogden, T. R. (2016). Methods for scalar-on-function regression. *International Statistical Review.* 2, 12

Robert, C. (1992). *La statistique Bayésienne.* Economica, Paris. 53

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer-Verlag New York. 18, 19

Robert, C. P. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer-Verlag New York. 22, 72

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367. 72

Rockova, V. and Lesaffre, E. (2014). Incorporating grouping information in bayesian variable selection with applications in genomics. *Bayesian Analysis*, 9. 54

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, 96(1):149–162. 12

Rudin, W. (1986). *Real and complex analysis*. McGraw-Hill Inc, New York, 3rd edition. 44

Scheipl, F., Staicu, A.-M., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501. 110

Schwartz, L. (1965). On bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26. 8, 94

Sen, P. K. and Singer, J. M. (1994). *Large sample methods in statistics: an introduction with applications*, volume 25. CRC Press. 102

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics*, pages 687–714. 94

Shi, J., Wang, B., Murray-Smith, R., and Titterington, D. (2007). Gaussian process functional regression modeling for batch data. *Biometrics*, 63(3):714–723. 12

Splivallo, R., Rittersma, R., Valdez, N., Chevalier, G., Molinier, V., Wipf, D., and Karlovsky, P. (2012). Is climate change altering the geographic distribution of truffles? . *Frontiers in Ecology and the Environment*, 10(9):461–462. 35

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418. 23

Stingo, F. and Vannucci, M. (2010). Variable selection for discriminant analysis with markov random field priors for the analysis of microarray data. *Bioinformatics*, 27. 54

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108. 13, 24

Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science*, 185:1124–1131. 55

Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):43. 2

Usset, J., Staicu, A.-M., and Maity, A. (2016). Interaction models for functional regression. *Computational statistics & data analysis*, 94:317–329. 110

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. 94

Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.*, 6:142–228. 70, 71, 89

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601. 9, 94, 105, 111

Walker, S. G. (2004). Modern bayesian asymptotics. *Statistical Science*, pages 111–117. 94

Wang, J. L., Chiou, J. M., and Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295. 2

Wang, X., Nan, B., Zhu, J., and Koeppe, R. (2014). Regularized 3d functional regression for brain image data via haar wavelets. *The annals of applied statistics*, 8(2):1045. 110

Wang, X., Ray, S., and Mallick, B. K. (2007). Bayesian curve classification using wavelets. *Journal of the American Statistical Association*, 102(479):962–973. 4

Wang, X., van Eeden, C., and Zidek, J. V. (2004). Asymptotic properties of maximum weighted likelihood estimators. *Journal of Statistical Planning and Inference*, 119(1):37 – 54. 60

Wasserman, L. (1998). Asymptotic properties of nonparametric bayesian procedures. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 293–304. Springer. 94

Winkler, R. (1967). The Assessment of Prior Distributions in Bayesian Analysis. *Journal of the American Statistical Association*, 62. 55

Winkler, R. (1968). The consensus of subjective probability distributions. *Management Science*, 15(2):B61–B75. 62

Winkler, R., Hora, S., and Baca, R. (1992). The quality of experts' probabilities obtained through formal elicitation techniques. *Center for Nuclear Waste Regulatory Analyses*. 55

Xiang, F. and Walker, S. G. (2013). Bayesian consistency for regression models under a supremum distance. *Journal of Statistical Planning and Inference*, 143(3):468–478. 94, 111

Yang, J., Cox, D. D., Lee, J. S., Ren, P., and Choi, T. (2017). Efficient bayesian hierarchical functional data analysis with basis function approximations using gaussian–wishart processes. *Biometrics*. 12

Yang, J., Zhu, H., Choi, T., Cox, D. D., et al. (2016). Smoothing and mean–covariance estimation of functional data with a bayesian hierarchical model. *Bayesian Analysis*, 11(3):649–670. 12

Yang, W.-H., Wikle, C. K., Holan, S. H., and Wildhaber, M. L. (2013). Ecological prediction with nonlinear multivariate time-frequency functional data models. *Journal of agricultural, biological, and environmental statistics*, 18(3):450–474. 2, 110

Yao, F., Fu, Y., and Lee, T. C. (2010). Functional mixture regression. *Biostatistics*, 12(2):341–353. 2

Yao, F. and Müller, H.-G. (2010). Functional quadratic regression. *Biometrika*, 97(1):49–64. 110

Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685. 23

Yao, F., Müller, H.-G., Wang, J.-L., et al. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903. 2

Yuan, M. and Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444. 12

Zellner, A. (1972). On assessing informative prior distributions for regression coefficients. 55

Zellner, A. (1986). *Bayesian inference and decision techniques - essays in honour of Bruno De Finetti*, chapter On assessing prior distributions and Bayesian regression analysis with g-prior distributions, pages 233–243. Elsevier Science Ltd, Amsterdam. 89

Zhang, T., Zhang, Q., and Wang, Q. (2014). Model detection for functional polynomial regression. *Computational Statistics & Data Analysis*, 70:183–197. 110

Zhao, Y., Ogden, T., and Reiss, P. (2012). Wavelet-Based LASSO in Functional Linear Regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617. 12

Zhou, J., Wang, N.-Y., and Wang, N. (2013). Functional Linear Model with Zero-Value Coefficient Function at Sub-Regions. *Statistica Sinica*, 23(1):25–50. 4, 13

Zhu, H., Yao, F., and Zhang, H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society Series B*, 76(3):581–603. 12

Résumé : Un outil fondamental en statistique est le modèle de régression linéaire. Lorsqu'une des covariables est une fonction, on fait face à un problème de statistique en grande dimension. Pour conduire l'inférence dans cette situation, le modèle doit être parcimonieux, par exemple en projetant la covariable fonctionnelle dans des espaces de plus petites dimensions.
Dans cette thèse, nous proposons une approche bayésienne nommée Bliss pour ajuster le modèle de régression linéaire fonctionnel. Notre modèle, plus précisément la distribution *a priori*, suppose que la fonction coefficient est une fonction en escalier. A partir de la distribution *a posteriori*, nous définissons plusieurs estimateurs bayésiens, à choisir suivant le contexte : un estimateur du support et deux estimateurs de la fonction coefficient, un lisse et un estimateur constant par morceaux. A titre d'exemple, nous considérons un problème de prédiction de la production de truffes noires du Périgord en fonction d'une covariable fonctionnelle représentant l'évolution des précipitations au cours du temps. En termes d'impact sur les productions, la méthode Bliss dégage alors deux périodes de temps importantes pour le développement de la truffe.
Un autre atout du paradigme bayésien est de pouvoir inclure de l'information dans la loi *a priori*, par exemple l'expertise des trufficulteurs et des biologistes sur le développement de la truffe. Dans ce but, nous proposons deux variantes de la méthode Bliss pour prendre en compte ces avis. La première variante récolte de manière indirecte l'avis des experts en leur proposant de construire des données fictives. La loi *a priori* correspond alors à la distribution *a posteriori* sachant ces pseudo-données. En outre, un système de poids relativise l'impact de chaque expert en prenant en compte leurs dépendances respectives. La seconde variante récolte explicitement l'avis des experts sur les périodes de temps les plus influentes sur la production et si cet impact est positif ou négatif. La construction de la loi *a priori* repose alors sur une pénalisation des fonctions coefficient en contradiction avec ces avis d'experts.
Enfin, ces travaux de thèse s'attachent à l'analyse et la compréhension du comportement de la méthode Bliss. La validité de l'approche est justifiée par une étude asymptotique de la distribution *a posteriori*. Nous avons construit un jeu d'hypothèses spécifiques au modèle Bliss, pour écrire une démonstration efficace d'un théorème de Wald. Une des difficultés est la mauvaise spécification du modèle Bliss, dans le sens où la vraie fonction coefficient n'est sûrement pas une fonction en escalier. Nous montrons que la loi *a posteriori* se concentre autour d'une fonction coefficient en escalier, obtenue par projection au sens de la divergence de Kullback-Leibler de la vraie fonction coefficient sur un ensemble de fonctions en escalier. Nous caractérisons cette fonction en escalier à partir du *design* et de la vraie fonction coefficient.

Mots clés : Statistique bayésienne, régression linéaire fonctionnelle, grande dimension, parcimonie, élicitation, information *a priori*, pseudo-données, pénalisation, asymptotique, consistance de la distribution *a posteriori*, modèle mal spécifié

Abstract : The linear regression model is a common tool for a statistician. If a covariable is a curve, we tackle a high-dimensional issue. In this case, sparse models lead to successful inference, for instance by expanding the functional covariate on a smaller dimensional space.
In this thesis, we propose a Bayesian approach, named Bliss, to fit the functional linear regression model. The Bliss model supposes, through the prior, that the coefficient function is a step function. From the posterior, we propose several estimators to be used depending on the context: an estimator of the support and two estimators of the coefficient function: a smooth one and a stepwise one. To illustrate this, we explain the black Périgord truffle yield with the rainfall during the truffle life cycle. The Bliss method succeeds in selecting two relevant periods for truffle development.
As another feature of the Bayesian paradigm, the prior distribution enables the integration of preliminary judgments in the statistical inference. For instance, the biologists' knowledge about the truffles growth is relevant to inform the Bliss model. To this end, we propose two modifications of the Bliss model to take into account preliminary judgments. First, we indirectly collect preliminary judgments using pseudo data provided by experts. The prior distribution proposed corresponds to the posterior distribution given the experts' pseudo data. Futhermore, the effect of each expert and their correlations are controlled with weighting. Secondly, we collect experts' judgments about the most influential periods effecting the truffle yield and if the effect is positive or negative. The prior distribution proposed relies on a penalization of coefficient functions which do not conform to these judgments.
Lastly, the asymptotic behavior of the Bliss method is studied. We validate the proposed approach by showing the posterior consistency of the Bliss model. Using model-specific assumptions, efficient proof of the Wald theorem is given. The main difficulty is the misspecification of the model since the true coefficient function is surely not a step function. We show that the posterior distribution contracts on a step function which is the Kullback-Leibler projection of the true coefficient function on a set of step functions. This step function is derived from the true parameter and the design.

Keywords : Bayesian statistics, functional linear regression, high-dimensional statistics, sparsity, elicitation, prior information, pseudo data, penalization, asymptotic properties, posterior consistency, misspecification