Consistance de la loi a posteriori pour un modèle mal spécifié de régression sur données fonctionnelles

Paul-Marie Grollemund [1] & Christophe Abraham [2]

[1] *IMAG UMR 5149, Université de Montpellier, CNRS, Place E. Bataillon, 34095 Montpellier CEDEX, France*

[2] *MISTEA UMR 729, Montpellier SupAgro, INRA, CNRS, Univ Montpellier, Place Pierre Viala, 34060 Montpellier CEDEX, France*

Résumé. Une première validation importante pour un modèle bayésien est l'établissement de la consistance du modèle. Autrement dit, on s'intéresse au comportement de la distribution *a posteriori* lorsque la taille de l'échantillon tend vers l'infini. Cette étude suppose qu'il existe *un vrai paramètre* θ_*, et on souhaite établir sous quelles conditions la distribution *a posteriori* se concentre autour de ce *vrai* paramètre. De nombreux travaux donnent des résultats généraux pour établir ce genre de consistance, ainsi que des résultats plus fins, pour une large gamme de modèles. Pendant cet exposé, nous présenterons un résultat, basé sur un jeu d'hypothèses simples et interprétables, pour établir la consistance d'un modèle de régression linéaire appliqué à des données fonctionnelles. La contribution principale de ce travail est d'adapter l'approche de Wald au cas d'un modèle de régression mal-spécifié.

Mots-clés. statistique bayésienne, consistance, modèle mal spécifié, régression linéaire, données fonctionnelles

Abstract. A major property of a Bayesian model is the posterior consistency, *i.e.* the posterior distribution behavior when the sample size increases. In order to show this property, it is required to assume the existence of a *true* parameter and the aim is to determine the necessary assumptions in such a way that the posterior contracts around the *true* parameter. Many authors give overall results consistency for a range of models. During the talk, we give a result based on simple and interpretable assumptions to show the consistency of the Functional Linear Regression model. The main contribution is to adapt the approach of Wald in the case of a misspecified regression model.

Keywords. Bayesian statitics, consistency, misspecified model, Linear Regression, functional data

1 Introduction

Notons $y_1, \dots, y_n$ des données réelles et $x_1(\cdot), \dots, x_n(\cdot)$ des courbes. Nous nous intéressons dans ce qui suit au modèle de régression linéaire sur données fonctionnelles défini, pour

le paramètre $\theta = (\mu, \beta(\cdot), \sigma^2)$ et pour tout $i = 1, \ldots, n$ par :

$$y_i | x_i(\cdot), \mu, \beta(\cdot), \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}(r(x_i)\,,\,\sigma^2) = P_{\theta,i}, \quad \text{où } r(x_i) = \mu + \int_0^1 x_i(t)\beta(t)\,\mathrm{d}t \tag{1}$$

où μ et σ^2 sont des réels et $\beta(\cdot)$ est une fonction. La principale difficulté pour ce modèle réside dans l'estimation de la fonction coefficient et on pourra consulter Reiss et al. (2016) pour une récente revue des méthodes proposées. Les approches standards consistent à écrire les fonctions $x_i(\cdot)$ et $\beta(\cdot)$ dans une base finie de fonctions afin de se ramener à une écriture plus simple. Différentes extensions de ce modèle on été envisagées, comme le modèle de régression généralisé, le modèle mixte, ou encore des extensions prenant en compte des interactions entre covariables fonctionnelles, des données fonctionnelles en 2D ou en 3D, *etc.* Lorsque l'objectif est de comprendre le lien linéaire entre les données réelles y_i et les courbes $x_i(\cdot)$, une quantité importante à estimer est le support de la fonction coefficient $\beta(\cdot)$. Plusieurs approches sont possibles pour estimer ce support. Par exemple, James et al. (2009) et Zhou et al. (2013) imposent une forte régularité sur la fonction coefficient, ce qui revient à estimer une version parcimonieuse de $\beta(\cdot)$, alors que Picheny et al. (2016); Park et al. (2016) proposent directement des estimateurs fréquentistes du support. Dans un cadre bayésien, Grollemund et al. (2018) propose d'écrire la fonction $\beta(\cdot)$ comme une fonction constante par morceaux de la manière suivante :

$$\beta(\cdot) \in \mathcal{E}_K = \left\{ \beta(t) = \sum_{k=1}^{K} b_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}_{\mathcal{I}_k}(t) \quad \text{où les } \mathcal{I}_k \subset [0,1] \right\}. \tag{2}$$

Les auteurs établissent un modèle bayésien dont la loi *a priori* charge l'ensemble de ces fonctions en escalier en favorisant celles dont le support (union des intervalles $\mathcal{I}_k$) ne recouvre pas $[0,1]$. En considérant une fonction de coût particulière, ils introduisent un estimateur bayésien du support ainsi que deux estimateurs bayésiens de la fonction coefficient. Le modèle ainsi proposé dépend d'un hyperparamètre important : K, le nombre d'escaliers de la fonction coefficient. Pour calibrer ce paramètre, les auteurs proposent de recourir à une procédure de choix de modèle en utilisant le critère d'information bayésien (BIC). Au-delà de cette manière de calibrer K, on peut se demander ce qu'il advient si la vraie fonction coefficient est en escalier avec un nombre d'escaliers K_* inférieur à K, autrement dit si $\beta_*(\cdot) \in \mathcal{E}_{K_*}$ avec $K_* \leq K$. Selon (2), les intervalles $\mathcal{I}_k$, peuvent se chevaucher, si bien qu'au moins intuitivement, il semble que si n est suffisamment grand, la loi *a posteriori* de $\beta(\cdot)$ se concentre autour de $\beta_*(\cdot)$ dès que $K_* \leq K$. Dans ce cas là, le problème du choix de K se résout en pratique en fixant K suffisamment grand. Inversement, si $K_* > K$ on peut se demander si la loi *a posteriori* se concentre encore et si oui, autour de quelle valeur ? Plus généralement pour un modèle mal spécifié où la vraie fonction $\beta_*(\cdot)$ n'est pas en escalier, ce qui est sûrement le cas en pratique, autour de quelle valeur la distribution *a posteriori* se concentre-t-elle ? Ainsi, par la suite, on suppose que le modèle est mal spécifié : on considère que les données sont générées selon

(1) avec $\beta_*(\cdot) \in L^2([0,1])$ qui n'est donc pas nécessairement une fonction en escalier alors que la loi a priori ne charge que les fonctions en escaliers de $\mathcal{E}_K$.

Il est important de noter que la consistance, c'est-à-dire la concentration de la loi *a posteriori* vers la vraie loi des observations, ne va pas de soi en général. Freedman (1963) propose un exemple célèbre dans lequel, bien que la loi *a priori* charge tout voisinage ouvert de la vraie loi des observations, la loi *a posteriori* se concentre autour d'une autre loi. Le lecteur pourra consulter Diaconis and Freedman (1986) pour une revue de la consistance dans un cadre bien spécifié. Dans un cadre mal spécifié, c'est-à-dire lorsque la vraie loi des observations n'est pas dans le support de la loi *a priori*, on s'attend généralement à ce que la loi *a posteriori* se concentre autour d'une "pseudo-vraie" loi, c'est-à-dire d'un élément du support de la loi *a priori* le plus proche de la vraie loi des observations. Cependant, là encore, Grünwald et al. (2017) propose un exemple d'inconsistance typique du cadre mal spécifié. La consistance de la distribution *a posteriori* est un point important à étudier puisqu'il existe des cas de modèles mal spécifiés pour lesquels la distribution *a posteriori* est inconsistante. Des cas d'inconsistance ont été par exemple donnés par Stone (1976) ou Freedman and Diaconis (1983); Diaconis and Freedman (1986), et pour un exemple plus récent, on pourra consulter Grünwald et al. (2017) qui montre l'inconsistance pour un modèle de régression supposant à tort l'homoscédasticité.
Pour montrer la consistance de la distribution *a posteriori*, principalement deux démarches ont été envisagées. La première est de celle Wald (1949), basée sur une étude du maximum de vraisemblance. La seconde est l'approche de Schwartz (1965) qui donne la consistance sous les hypothèses 1) que la loi *a priori* charge les voisinages de Kullback-Leibler du *vrai* paramètre et 2) qu'il existe une suite de fonctions de test consistante. Cette dernière démarche est la plus utilisée dans la littérature, puisqu'elle s'applique plus efficacement au cadre non-paramétrique (voir Ghosh and Ramamoorthi, 2003). À partir de ce résultat, de nombreux travaux se sont intéressés à des extensions ou à des résultats plus fins et on pourra consulter Kleijn (2016) pour un résumé de ces variantes.

Parmi les variantes étudiées, une nous intéresse ici en particulier : la consistance lorsque le modèle est mal spécifié. En adaptant les travaux de Schwartz (1965), Kleijn and van der Vaart (2006) propose un cadre général pour établir la consistance de modèles non-paramétriques mal spécifiés, au prix d'hypothèses assez techniques. Dans notre contexte, nous travaillons avec un modèle de régression mal spécifié et pour établir la consistance du modèle en dégageant des hypothèses simples et interprétables, nous adaptons la démarche de Wald (1949) au cadre mal spécifié. Le reste de ce résumé comprend la section 2 qui introduit quelques notations ainsi que les hypothèses nécessaires à l'établissement des résultats donnés en section 3.

2 Notations et hypothèses

Supposons qu'il existe un *vrai* paramètre $\theta_* = (\mu_*, \beta_*(\cdot), \sigma_*^2)$ et qu'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 $x_i(\cdot) \in L^2([0,1])$ telles que :

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

Les distributions produits du modèle sont notées P_θ^n et P_θ^∞ ($P_{\theta_*}^n$ et $P_{\theta_*}^\infty$ pour le *vrai* modèle). 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 de même pour $P_{\theta_*}^\infty$. Le paramètre θ appartient à $\Theta \subset \mathbb{R} \times \mathcal{E}_K \times \mathbb{R}_+^*$, ensemble sur lequel on définit la norme $\|\theta\| = |1/\sigma^2| + \|\beta(\cdot)\|_{L^2} + |\mu|$, afin de travailler efficacement avec la vraisemblance du modèle gaussien.

La loi *a priori* sur Θ est notée par Π et nous notons par Π_n la distribution *a posteriori* sur Θ sachant les données $\{y_i, x_i\}$ pour $i = 1, \ldots, n$.

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 $f^{\otimes 2} = f \otimes f$. De plus, on note par $\bar{x}_n(\cdot) = \frac{1}{n}\sum_{i=1}^n x_i(\cdot)$ la moyenne empirique de $x_1(\cdot), \ldots, x_n(\cdot)$ et $\overline{x_n^2}(\cdot,\cdot) = \frac{1}{n}\sum_{i=1}^n x_i^{\otimes 2}(\cdot,\cdot)$.

Contrairement aux cas où le modèle est bien spécifié, la distribution *a posteriori* ne peut pas se concentrer autour du vrai paramètre θ_* puisque ce dernier n'est pas dans son support. Dans cette section, nous donnons des hypothèses, que nous discuterons pendant la présentation, afin d'établir que la distribution *a posteriori* du modèle Bliss se concentre autour d'un paramètre $\theta_0 \in \Theta$. Les deux premières hypothèses sont des hypothèses raisonnables sur le *design*, considéré non-aléatoire dans notre contexte. La première impose que les courbes $x_i(t)$ soient bornées, uniformément en i et en t. La deuxième hypothèse demande une convergence simple de $\bar{x}_n(\cdot)$ et $\overline{x_n^2}(\cdot,\cdot)$ ce qui revient à imposer une certaine régularité sur le *design*.

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

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

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

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

La troisième hypothèse permet de définir une solution $\beta_0(\cdot)$ comme le minimum d'un critère dépendant du *design*.

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

$$F(\beta) = \iint (\beta_* - \beta)^{\otimes 2}(t,t') \left[c(t,t') - e^{\otimes 2}(t,t')\right] \mathrm{d}t\, \mathrm{d}t'. \tag{3}$$

La dernière hypothèse, issue des travaux de Wald (1949), peut être relaxée et permet d'avoir une démonstration simple du résultat.

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

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

3 Résultats

À partir des hypothèses 1 et 2 sur le *design*, on montre avec la proposition 1 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 1. *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 $\mathcal{E}_K$ ce qui nous permet de montrer avec la proposition 2 que θ_0 s'exprime à partir du *design* et de θ_*.

Proposition 2. *Sous les hypothèses 1 à 4,* $\theta_0 = (\mu_0, \beta_0(\cdot), \sigma_0^2)$ *où* $\beta_0(\cdot)$ *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 $\beta_*(\cdot)$ est une fonction en escalier, avec K escaliers ou moins, alors θ_0 coïncide avec le *vrai* paramètre θ_*. De plus, on peut montrer 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.

On obtient alors la consistance de la distribution *a posteriori* en $\theta_0 \in \Theta$ avec le théorème 1.

Théorème 1. *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* $n \to +\infty$*.*

Notre approche pour prouver le Théorème 1 consiste à adapter la preuve de Wald (1949) à notre cadre où le modèle est mal spécifié. Ainsi, à partir de l'expression de la distribution *a posteriori* donnée par

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

on fait apparaître au numérateur et au dénominateur, la moyenne du rapport de log-vraisemblance $n^{-1}\sum_{i=1}^n -\log p_{\theta,i}/p_{\theta_0,i}$. On montre avec une loi forte des grands nombres que cette moyenne tends uniformément en θ vers

$$-P^\infty_{\theta_*} \log \frac{p^\infty_\theta}{p^\infty_{\theta_0}} = \bar{K}(\theta).$$

Ce dernier terme s'apparente à la divergence de Kullback-Leibler et on détermine grâce à la proposition 2 qu'il atteint son minimum en θ_0. Puis, on détermine deux voisinages de θ_0 tels que $\bar{K}(\theta)$ soit plus petit à l'intérieur de l'un qu'à l'extérieur de l'autre, ce qui permet de contrôler le numérateur et le dénominateur afin de majorer $\Pi_n(U)$ par un terme qui tend exponentiellement vers 0. On montre ensuite que la loi *a priori* charge les voisinages de θ_0 ce qui permet de conclure la démonstration.

References

Diaconis, P. and Freedman, D. (1986). On inconsistent bayes estimates of location. The Annals of Statistics, pages 68–87.

Freedman, D. and Diaconis, P. (1983). On inconsistent bayes estimates in the discrete case. The Annals of Statistics, pages 1109–1118.

Freedman, D. A. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. Ann. Math. Statist., 34(4):1386–1403.

Ghosh, J. and Ramamoorthi, R. (2003). Bayesian Nonparametrics. Springer New York, New York, NY.

Grollemund, P.-M., Abraham, C., Baragatti, M., and Pudlo, P. (2018). Bayesian functional linear regression with sparse step functions. Statistics, Methodology arXiv:1604.08403.

Grünwald, P., van Ommen, T., et al. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. Bayesian Analysis, 12(4):1069–1103.

James, G., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. The Annals of Statistics, 37(5A):2083–2108.

Kleijn, B. (2016). On the frequentist validity of bayesian limits. arXiv preprint arXiv:1611.08444.

Kleijn, B. J. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional bayesian statistics. The Annals of Statistics, pages 837–877.

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.

Picheny, V., Servien, R., and Villa-Vialaneix, N. (2016). Interpretable sparse sir for functional data. arXiv preprint arXiv:1606.00614.

Reiss, P., Goldsmith, J., Shang, H., and Ogden, T. R. (2016). Methods for scalar-on-function regression. International Statistical Review.

Schwartz, L. (1965). On bayes procedures. Probability Theory and Related Fields, 4(1):10–26.

Stone, M. (1976). Strong inconsistency from uniform priors. Journal of the American Statistical Association, 71(353):114–116.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. The Annals of Mathematical Statistics, 20(4):595–601.

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.