Le théorème de Bernstein-von Mises pour la régression gaussienne sous un nombre croissant de régresseurs

Dominique Bontemps

Institut de Mathématiques de Toulouse

Séminaire de l'unité BIA
27 janvier 2012

Outline

Bayesian paradigm

If $(P_\theta)_{\theta\in\Theta}$ is a statistical model, a Bayesian puts a prior distribution W on θ. Given a risk function L, an estimator of $g(\theta)$

$$\hat{g}^W = \arg\min_{\delta} \int_\Theta L(g(\theta), \delta) W(d\theta|\boldsymbol{X})$$

is built on the basis of the posterior distribution given the data

$$W(d\theta|\boldsymbol{X}) = \frac{dP_\theta(\boldsymbol{X})\, dW(\theta)}{\int_\Theta dP_\nu(\boldsymbol{X})\, dW(\nu)}.$$

Smooth parametric models

Consider $(P_\theta)_{\theta\in\Theta}$ a smooth statistical model, with Θ a domain in $\mathbb{R}^k$ and the log-likelihood

$$\ell_\theta(\boldsymbol{X}) = \log\left(dP_\theta(\boldsymbol{X})\right)$$

be C^2 for $\theta \in \Theta$. The Fisher Information matrix is defined as

$$I_\theta = E_\theta\left[\dot\ell_\theta(\boldsymbol{X})\dot\ell_\theta^T(\boldsymbol{X})\right].$$

Suppose the model to be identifiable, $\boldsymbol{X} \sim P_0 = P_{\theta_0}$, and I_{θ_0} to be invertible. With high probability the Maximum Likelihood Estimator $\widehat\theta^{MLE}$ exists and the log-likelihood admits a quadratic development at the neighborhood of θ_0:

$$\ell_{\theta_0+h} = \ell_{\theta_0} + h^T I_{\theta_0}\left(\widehat\theta^{MLE} - \theta_0\right) - \frac{1}{2}h^T I_{\theta_0} h + o_{P_{\theta_0}}(1).$$

Suppose $\boldsymbol{X} \sim P_0$ and $g(P_0)$ is a quantity of interest,

- Is $W(dg(P_\theta)|\boldsymbol{X})$ concentrated near $g(P_0)$?
- Is $W(dg(P_\theta)|\boldsymbol{X})$ approximately Gaussian?

What if P_0 is outside the model?

The i.i.d. parametric Bernstein-von Mises Theorem

- A parametric model $(P_\theta)_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^k$, identifiable, q.m.d. at θ_0 in the interior of Θ, with invertible Fisher Information I_{θ_0};
- $X_{1:n} = X_1, \dots, X_n$ i.i.d. following $P_0 = P_{\theta_0}$;
- $W(d\theta)$ a prior on Θ, with density w continuous and positive at θ_0;

Then the MLE $\widehat{\theta}^{MLE}$ exists with probability going to 1, it converges in distribution towards $\mathcal{N}\left(\theta_0, \frac{1}{n} I_{\theta_0}^{-1}\right)$, and

$$E \left\| W(d\theta|X_{1:n}) - \mathcal{N}\left(\widehat{\theta}^{MLE}, \frac{1}{n} I_{\theta_0}^{-1}\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Comments

$W(d\theta|X_{1:n})$ is approximately $\mathcal{N}\left(\widehat{\theta}^{MLE}, \frac{1}{n} I_{\theta_0}^{-1}\right)$;

$\widehat{\theta}^{MLE}$ is approximately $\mathcal{N}\left(\theta_0, \frac{1}{n} I_{\theta_0}^{-1}\right)$:

(Bayesian) credibility intervals and (frequentist) confidence intervals are asymptotically the same.

Comments

$W(d\theta|X_{1:n})$ is approximately $\mathcal{N}\left(\widehat{\theta}^{MLE}, \frac{1}{n} I_{\theta_0}^{-1}\right)$;
$\widehat{\theta}^{MLE}$ is approximately $\mathcal{N}\left(\theta_0, \frac{1}{n} I_{\theta_0}^{-1}\right)$:
(Bayesian) credibility intervals and (frequentist) confidence intervals are asymptotically the same.

Motivation from Information Theory:

$$\inf_{Q^n} \sup_{\theta \in \Theta} D(P_\theta^n; Q^n) = \sup_W \inf_{Q^n} \int_\Theta D(P_\theta^n; Q^n)\, W(d\theta).$$

The infimum on the right side is achieved by the Bayes mixture $M_W^n(x_{1:n}) = \int_\Theta P_\theta^n(x_{1:n})\, W(d\theta)$, and

$$D\left(P_\theta^n; M_W^n\right) = E_{P_\theta^n}\left[\log \frac{W(d\theta|X_{1:n})}{W(d\theta)}\right]$$

Previous semiparametric and nonparametric results

- Many results about the posterior convergence rates in various nonparametric settings;
- Some semiparametric Bernstein-von Mises theorems: Kim and Lee (2004), Kim (2006), Shen (2002), Castillo (2009), Rivoirard and Rousseau (2009);
- Nonparametric Bernstein-von Mises theorems in increasing dimension settings: Ghosal (1999), Ghosal (2000), Boucheron and Gassiat (2009), Clarke and Ghosal (2010).

The Regression model with Gaussian noise

- The observation $Y = (Y_1, \ldots, Y_n)$ is a Gaussian random vector
$$Y = F_0 + \varepsilon$$
where $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ and $F_0 \in \mathbb{R}^n$.
- A (misspecified) model $P_\theta = \mathcal{N}(\Phi\theta, \sigma_n^2 I_n)$, where Φ is a $n \times k_n$ matrix whose columns $\phi_1, \ldots, \phi_{k_n}$ are linearly independent regressors. k_n grows with n.
Let $\langle\phi\rangle$ be the linear span of the regressors, and $\Sigma = \Phi(\Phi^T\Phi)^{-1}\Phi^T$ the matrix of the orthogonal projection on $\langle\phi\rangle$.
- A prior distribution $W(dF)$ on $\langle\phi\rangle$, induced by the distribution $\widetilde{W}(d\theta) = w(\theta)d\theta$ on $\mathbb{R}^{k_n}$ by the map $F = \Phi\theta$.
- The MLE is $Y_{\langle\phi\rangle} = \Phi\theta_Y = \Sigma Y$. Let $F_{\langle\phi\rangle} = \Phi\theta_0 = \Sigma F_0$. Then
$$Y_{\langle\phi\rangle} \sim \mathcal{N}\left(F_{\langle\phi\rangle}, \sigma_n^2\Sigma\right).$$

Example: The Gaussian sequence model

$$Y_j = \alpha_j^0 + \frac{1}{\sqrt{n}}\xi_j, \qquad j \geq 1$$

where $\xi_j, j \geq 1$ are i.i.d. $\mathcal{N}(0,1)$.

- This is linked to the white noise model.
- We project on the first k_n coordinates, so $F_{\langle\phi\rangle} = (\alpha_j^0)_{1\leq j\leq k_n}$ and the MLE $Y_{\langle\phi\rangle} = (Y_j)_{1\leq j\leq k_n}$.
- α^0 is supposed to be in a Sobolev class: for some $\beta > 0$, $\sum_{j=1}^{\infty} |\alpha_j^0|^2 j^{2\beta} < \infty$.

Example: Regression of a C^α function

Let $\alpha > 0$, and α_0 be the integer part of α. We define a seminorm on $C^\alpha[0,1]$

$$\|f\|_\alpha = \sup_{x \neq x'} \frac{\left|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(x')\right|}{|x - x'|^{\alpha-\alpha_0}}.$$

Consider a design $\left(x_i^{(n)}\right)_{n\geq 1, 1\leq i \leq n}$, not necessarily uniform. We observe the vector $\left(f(x_i^{(n)}) + \varepsilon_i\right)_{1\leq i\leq n}$, and want to retrieve f. Here $F_0 = \left(f(x_i^{(n)})\right)_{1\leq i \leq n}$ and $\sigma_n = \sigma$ is constant.
Regressors: fix an integer $q \geq \alpha$, and let $K = k_n + 1 - q$. Let $(B_j)_{1\leq j\leq k_n}$ be the B-splines of order q on the regular partition of $[0,1]$ into K subintervals. Then $\phi_j = \left(B_j(x_i^{(n)})\right)_{1\leq i\leq n}$ for $1 \leq j \leq k_n$.

Example: Regression of a C^α function

For $\theta \in \mathbb{R}^{k_n}$, let $f_\theta = \sum_{j=1}^{k_n} \theta_j B_j$.

Approximation property of the B-splines

For any $\alpha > 0$, there exist $C_\alpha > 0$ such that, if $f \in C^\alpha[0, 1]$, there exists $\theta^\infty \in \mathbb{R}^{k_n}$ verifying

$$\|f - f_{\theta^\infty}\|_\infty \leq C_\alpha k_n^{-\alpha} \|f\|_\alpha.$$

Example: Regression of a C^α function

For $\theta \in \mathbb{R}^{k_n}$, let $f_\theta = \sum_{j=1}^{k_n} \theta_j B_j$.
Approximation property of the B-splines
For any $\alpha > 0$, there exist $C_\alpha > 0$ such that, if $f \in C^\alpha[0,1]$, there exists $\theta^\infty \in \mathbb{R}^{k_n}$ verifying

$$\|f - f_{\theta^\infty}\|_\infty \leq C_\alpha k_n^{-\alpha} \|f\|_\alpha.$$

A norm $\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n |f(x_i)|^2}$ is associated to the design $\left(x_i^{(n)}\right)_{n\geq 1, 1\leq i\leq n}$. The design is supposed to be sufficiently regular, so that there exist positive constants C_1 and C_2 such that, as n increases, whatever $\theta \in \mathbb{R}^{k_n}$,

$$C_1 \frac{n}{k_n} \|\theta\|^2 \leq \theta^T \Phi^T \Phi \theta \leq C_2 \frac{n}{k_n} \|\theta\|^2.$$

With an isotropic Gaussian prior

Theorem

Let $W = \mathcal{N}(0, \tau_n^2 \Sigma)$. *Assume that* $\sigma_n = o(\tau_n)$, $\|F_0\| = o(\tau_n^2/\sigma_n)$ *and* $k_n = o(\tau_n^4/\sigma_n^4)$. *Then*

$$E \left\| W(dF|Y) - \mathcal{N}\left(Y_{\langle\phi\rangle}, \sigma_n^2 \Sigma\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

With an isotropic Gaussian prior

Theorem

Let $W = \mathcal{N}(0, \tau_n^2 \Sigma)$. *Assume that* $\sigma_n = o(\tau_n)$, $\|F_0\| = o(\tau_n^2/\sigma_n)$ *and* $k_n = o(\tau_n^4/\sigma_n^4)$. *Then*

$$E \left\| W(dF|Y) - \mathcal{N}\left(Y_{\langle\phi\rangle}, \sigma_n^2 \Sigma\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Example: Regression of a bounded function f, with constant noise σ^2. The Bernstein-von Mises theorem holds as soon as $n^{1/4} = o(\tau_n)$.

Application to the Gaussian Sequence Model

Proposition

Suppose that $\sum_{j=1}^{k_n} (\theta_j^0)^2$ *is bounded. Let* $W = \mathcal{N}\left(0, \tau_n^2 I_{k_n}\right)$ *with* $n^{-1/4} = o(\tau_n)$*. Then, whatever* $k_n \leq n$*,*

$$E \left\| W(dF|Y) - \mathcal{N}\left(Y_{\langle\phi\rangle}, \frac{1}{n} I_{k_n}\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Let $\beta > 0$*, and suppose further that* $\sum_{j=1}^{\infty} |\alpha_j^0|^2 j^{2\beta} < \infty$*. Let* k_n *be of order* $n^{1/(1+2\beta)}$*.*
Then the convergence rate of F *towards* α^0 *is* $n^{-\beta/(1+2\beta)}$*: for every* $\lambda_n \to \infty$*,*

$$E\left[W\left(\|F - \alpha^0\| \geq \lambda_n n^{-\beta/(1+2\beta)} \,\middle|\, Y\right)\right] \to 0.$$

With a smooth prior

Theorem

Suppose that there exists a sequence $(M_n)_{n\geq 1}$ *such that*

1. $$\sup_{\|\Phi h\|^2 \leq \sigma_n^2 M_n, \|\Phi g\|^2 \leq \sigma_n^2 M_n} \frac{w(\theta_0 + h)}{w(\theta_0 + g)} \to 1 \text{ as } n \to \infty.$$
2. $k_n \ln k_n = o(M_n)$
3. $$\max\left(0, \ln\left(\frac{\sqrt{\det(\Phi^T \Phi)}}{\sigma_n^{k_n} w(\theta_0)}\right)\right) = o(M_n)$$

Then

$$E\left\|W(dF|Y) - \mathcal{N}\left(Y_{\langle\phi\rangle}, \sigma_n^2 \Sigma\right)\right\|_{TV} \to 0 \text{ as } n \to \infty.$$

Application to C^α functions

Proposition

Assume that f is bounded. Let $\widetilde{W} = \mathcal{N}\left(0, \tau_n^2 I_{k_n}\right)$ be the prior on the spline coefficients, with the sequence τ_n verifying $\frac{k_n^2 \ln n}{n} = o(\tau_n^2)$ and $\frac{k_n^3 \ln n}{n} = o(\tau_n^4)$. Then

$$E\left\|\widetilde{W}(d\theta|Y) - \mathcal{N}\left(\theta_Y, \sigma^2(\Phi^T\Phi)^{-1}\right)\right\|_{TV} \to 0.$$

Let $\alpha > 0$, and suppose further that f is C^α and k_n is of order $n^{1/(1+2\alpha)}$. Then the conditions reduce to $n^{\frac{2-2\alpha}{1+2\alpha}} \ln n = o(\tau_n^4)$ and, if this holds, the posterior concentrates at the minimax rate $n^{-\alpha/(1+2\alpha)}$ relative to $\|\cdot\|_n$: for every $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|f_\theta - f\|_n \geq \lambda_n n^{-\alpha/(1+2\alpha)} \middle| Y\right)\right] \to 0.$$

Linear functionals

Consider the estimation a linear functional GF_0 of F_0.

Corollary

Let $p \geq 1$ fixed, and G be a $\mathbb{R}^p \times \mathbb{R}^n$-matrix. Suppose that the conditions of either Theorem 1 or Theorem 2 are verified. Then

$$E \left\| W(d(GF)|Y) - \mathcal{N}\left(GY_{\langle\phi\rangle}, \sigma_n^2 G\Sigma G^T\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Further, the distribution of $GY_{\langle\phi\rangle}$ is $\mathcal{N}\left(GF_{\langle\phi\rangle}, \sigma_n^2 G\Sigma G^T\right)$.

Bias $GF_0 - GF_{\langle\phi\rangle}$?

Smooth functionals: conditions

Let $p \geq 1$ fixed, and $G : \mathbb{R}^n \mapsto \mathbb{R}^p$ be C^2. For any $F \in \langle \phi \rangle$ and $a > 0$, let

$$B_F(a) = \sup_{h \in \langle \phi \rangle : \|h\|^2 \leq \sigma_n^2 a} \sup_{0 \leq t \leq 1} \left\| D^2_{F+th} G(h, h) \right\|$$

and

$$\Gamma_F = \sigma_n^2 \dot{G}_F \Sigma \dot{G}_F^T .$$

Smooth functionals: conditions

Let $p \geq 1$ fixed, and $G : \mathbb{R}^n \mapsto \mathbb{R}^p$ be C^2. For any $F \in \langle \phi \rangle$ and $a > 0$, let

$$B_F(a) = \sup_{h \in \langle \phi \rangle : \|h\|^2 \leq \sigma_n^2 a} \sup_{0 \leq t \leq 1} \left\| D^2_{F+th} G(h,h) \right\|$$

and

$$\Gamma_F = \sigma_n^2 \dot{G}_F \Sigma \dot{G}_F^T .$$

Suppose that $\Gamma_{F_{\langle \phi \rangle}}$ is nonsingular, and that there exists a sequence $(M_n)_{n \geq 1}$ such that $k_n = o(M_n)$ and

$$B^2_{F_{\langle \phi \rangle}}(M_n) = o\left(\left\| \Gamma^{-1}_{F_{\langle \phi \rangle}} \right\|^{-1} \right) .$$

Suppose further that the conditions of either Theorem 1 or Theorem 2 (with the same sequence M_n) are verified.

Theorem

Then, for any $b \in \mathbb{R}^p$,

$$E\left[\sup_{I\in\mathcal{I}}\left|W\left(\frac{b^T\left(G(F)-G(Y_{\langle\phi\rangle})\right)}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}}\in I\,\middle|\,Y\right)-\psi(I)\right|\right]\to 0$$

where $\mathcal{I}$ is the collection of all intervals in $\mathbb{R}$, and for any $I\in\mathcal{I}$, $\psi(I)=P(Z\in I)$ if $Z\sim\mathcal{N}(0,1)$.
Under the same conditions,

$$\sup_{I\in\mathcal{I}}\left|P\left(\frac{b^T\left(G(Y_{\langle\phi\rangle})-G(F_{\langle\phi\rangle})\right)}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}}\in I\right)-\psi(I)\right|\to 0.$$

Proposition

Let $\beta > 1/2$ *and suppose that* $\sum_{j=1}^{\infty} |\alpha_j^0|^2 j^{2\beta} < \infty$. *Let* $W = \mathcal{N}\left(0, \tau_n^2 I_{k_n}\right)$ *with* $n^{-1/4} = o(\tau_n)$. *Then, for any choice of* k_n *such that* $k_n = o(\sqrt{n})$ *and* $\sqrt{n} = o(k_n^{2\beta})$,

$$E\left[\sup_{I\in\mathcal{I}}\left|W\left(\frac{\sqrt{n}\left(\|F\|^2 - \|Y_{\langle\phi\rangle}\|^2\right)}{2\|\alpha^0\|} \in I \,\middle|\, Y\right) - \psi(I)\right|\right] \to 0$$

and $\frac{\sqrt{n}\left(\|Y_{\langle\phi\rangle}\|^2 - \|F_{\langle\phi\rangle}\|^2\right)}{2\|\alpha^0\|} \xrightarrow{(d)} \mathcal{N}(0,1)$. *Further,*

$$\frac{\sqrt{n}\left(\|F_{\langle\phi\rangle}\|^2 - \|\alpha^0\|^2\right)}{2\|\alpha^0\|} = o(1).$$

In particular the choice $k_n = \sqrt{n/\ln n}$ *is adaptive in* β.

Merci pour votre attention.