Bernstein von Mises Theorems for Gaussian Regression with increasing number of regressors

Dominique Bontemps

The Annals of Statistics
2011, Vol. 39, No. 5, 2557–2584
DOI: 10.1214/11-AOS912

BERNSTEIN–VON MISES THEOREMS FOR GAUSSIAN REGRESSION WITH INCREASING NUMBER OF REGRESSORS

By Dominique Bontemps

Université Paris-Sud

This paper brings a contribution to the Bayesian theory of nonparametric and semiparametric estimation. We are interested in the asymptotic normality of the posterior distribution in Gaussian linear regression models when the number of regressors increases with the sample size. Two kinds of Bernstein–von Mises theorems are obtained in this framework: nonparametric theorems for the parameter itself, and semiparametric theorems for functionals of the parameter. We apply them to the Gaussian sequence model and to the regression of functions in Sobolev and C^α classes, in which we get the minimax convergence rates. Adaptivity is reached for the Bayesian estimators of functionals in our applications.

1. Introduction. To estimate a parameter of interest in a statistical model, a Bayesian puts a prior distribution on it and looks at the posterior distribution, given the observations. A Bernstein–von Mises theorem is a result giving conditions under which the posterior distribution is asymptotically normal, centered at the maximum likelihood estimator (MLE) of the model used, with a variance equal to the asymptotic frequentist variance of the MLE. Other centering can be used; see, for instance, van der Vaart (1998), page 144, after the proof of Lemma 10.3.

Such an asymptotic posterior normality is important because it allows the construction of approximate credible regions, based on the posterior distribution, which retain good frequentist properties. In particular, the Monte Carlo Markov chain algorithms (MCMC) make feasible the construction of Bayesian confidence regions in complex models, for which frequentist confidence regions are difficult to build; however, Bernstein–von Mises theorems are difficult to derive in complex models.

Received September 2010; revised June 2011.
AMS 2000 subject classifications. 62F15, 62J05, 62G20.
Key words and phrases. Nonparametric Bayesian statistics, semiparametric Bayesian statistics, Bernstein–von Mises theorem, posterior asymptotic normality, adaptive estimation.

Note that the Bernstein–von Mises theorem also has links with information theory [see Clarke and Barron (1990) and Clarke and Ghosal (2010)].

For parametric models, the Bernstein–von Mises theorem is a well-known result, for which we refer to van der Vaart (1998). In nonparametric models (where the parameter space is infinite-dimensional or growing) and semiparametric models (when the parameter of interest is a finite-dimensional functional of the complete infinite-dimensional parameter), there are still relatively few asymptotic normality results. Freedman (1999) gives negative results, and we recall some positive ones below. However, many recent papers deal with the convergence rate of posterior distributions in various settings, which is linked with the model complexity: we refer to Ghosal, Ghosh and van der Vaart (2000), Shen and Wasserman (2001) as early representatives of this school.

Nonparametric Bernstein–von Mises theorems have been developed for models based on a sieve approximation, where the dimension of the parameter grows with the sample size. In particular, two situations have been studied: regression models in Ghosal (1999); exponential models in Ghosal (2000), Clarke and Ghosal (2010) and Boucheron and Gassiat (2009) (this last one deals with the discrete case, when the observations follow some unknown infinite multinomial distribution).

In semiparametric frameworks the asymptotic normality has been obtained in several situations. Kim and Lee (2004) and Kim (2006) study the nonparametric right-censoring model and the proportional hazard model. Castillo (2010) obtains Bernstein–von Mises theorems for Gaussian process priors, in the semiparametric framework where the unknown quantity is (θ, f), with θ the parameter of interest and f an infinite-dimensional nuisance parameter. See also Shen (2002). Rivoirard and Rousseau (2009) obtain the Bernstein–von Mises theorem for linear functionals of the density of the observations, in the context of a sieve approximation: sequences of spaces with an increasing dimension k_n are used to approximate an infinite-dimensional space. These authors achieve also the frequentist minimax estimation rate for densities in specific regularity classes with a deterministic (nonadaptive) value of the dimension k_n.

Here we obtain nonparametric and semiparametric Bernstein–von Mises theorems in a Gaussian regression framework with an increasing number of regressors. We address two challenging problems. First, we try to understand better when the Bernstein–von Mises theorem holds and when it does not. In the latter case the Bayesian credible sets no longer preserve their frequentist asymptotic properties. Second, we look for adaptive Bayesian estimators in our semiparametric settings.

Our nonparametric results cover the case of a specific Gaussian prior, and the case of more generic smooth priors. They are said to be nonparametric because we use sieve priors, that is, the dimension of the parameter grows.

These results improve on the preceding ones by Ghosal (1999) which did not suppose the normality of the errors but imposed other conditions, in particular, on the growth rate of the number of regressors. We apply our results to the Gaussian sequence model, as well as to periodic Sobolev classes and to regularity classes $C^\alpha[0,1]$ in the context of the regression model (using, resp., trigonometric polynomials and splines as regressors). In all these situations we get the asymptotic normality of the posterior in addition to the minimax convergence rates, with appropriate (nonadaptive) choices of the prior. We also show that for some priors known to reach this convergence rate, the Bernstein–von Mises theorem does not hold.

We derive also semiparametric Bernstein–von Mises theorems for linear and nonlinear functionals of the parameter. The linear case is an immediate corollary of the nonparametric theorems and does not need any additional conditions. We apply these results to the periodic Sobolev classes to estimate a linear functional and the L^2 norm of the regression function f when it is smooth enough, and in both cases we are able to build an adaptive Bayesian estimator which achieves the minimax convergence rate in all classes of the collection, in addition to the asymptotic normality.

The paper is organized as follows. We present the framework in Section 2. Section 3 states the nonparametric Bernstein–von Mises theorems, for Gaussian and non-Gaussian priors. In Section 4 we derive the semiparametric Bernstein–von Mises theorems for linear and nonlinear functionals of the parameter. Then in Section 5 we give applications to the Gaussian sequence model, and to the regression of a function in a Sobolev and $C^\alpha[0,1]$ class. In Section 6 the nonparametric and semiparametric Bernstein–von Mises theorems are proved. The appendices contain various technical tools used in the main analysis; the appendices can be found in the supplemental article [Bontemps (2011)].

2. Framework. We consider a Gaussian linear regression framework. For any $n \geq 1$, our observation $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ is a Gaussian random vector

$$Y = F + \varepsilon, \tag{1}$$

where the vector of errors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \sim \mathcal{N}(0, \sigma_n^2 I_n)$, with I_n the $n \times n$ identity matrix, and the mean vector F belongs to $\mathbb{R}^n$. Note that the dimension of Y is the sample size n, and that σ_n^2 is known but may depend on n. Let F_0 be the true mean vector of Y with distribution $\mathcal{N}(F_0, \sigma_n^2 I_n)$. Probability expectations under F_0 are denoted P_{F_0} and E.

Let $\phi_1, \ldots, \phi_{k_n}$ a collection of k_n linearly independent regressors in $\mathbb{R}^n$, where $k_n \leq n$ grows with n. We gather these regressors in the $n \times k_n$-matrix Φ of rank k_n, and $\langle \phi \rangle = \{\Phi\theta : \theta = (\theta_1, \ldots, \theta_{k_n}) \in \mathbb{R}^{k_n}\}$ denotes their linear span. The Bernstein–von Mises theorems will be stated in association with $\langle \phi \rangle$, the

vector space of possible mean vectors in the model, which is possibly misspecified. We denote by P_θ the probability distribution of a random variable following $\mathcal{N}(\Phi\theta, \sigma_n^2 I_n)$ and E_θ the associated expectation.

As examples, we present three different settings, each with its own collection of regressors. In Section 5 the Bernstein–von Mises theorems are applied to each of these frameworks:

(1) *The Gaussian sequence model.* Our first application concerns the Gaussian sequence model, which is also equivalent to the white noise model [see Massart (2007), Chapter 4, e.g.]. We consider the infinite-dimensional setting

$$Y_j = \theta_j^0 + \frac{1}{\sqrt{n}}\xi_j, \qquad j \geq 1, \tag{2}$$

where the random variables $\xi_j, j \geq 1$ are independent and have distribution $\mathcal{N}(0,1)$. Projecting on the first k_n coordinates with $k_n \leq n$, we retrieve our model (1) with $\theta_0 = (\theta_j^0)_{1\leq j\leq k_n}$, $\sigma_n = 1/\sqrt{n}$ and $\Phi^T\Phi = I_{k_n}$.

(2) *Regression of a function in a Sobolev class.* Let $f : [0,1] \to \mathbb{R}$ be a function in $\mathbb{L}^2([0,1])$. We observe realizations of random variables

$$Y_i = f(i/n) + \varepsilon_i \tag{3}$$

for $1 \leq i \leq n$, where the errors ε_i are i.i.d. $\mathcal{N}(0, \sigma_n^2)$ and σ_n does not depend on n.

We denote by $(\varphi_j)_{j\geq 1}$ the Fourier basis

$$\begin{aligned} \varphi_1 &\equiv 1, \\ \varphi_{2m}(x) &= \sqrt{2}\cos(2\pi m x) \qquad \forall m \geq 1, \\ \varphi_{2m+1}(x) &= \sqrt{2}\sin(2\pi m x) \qquad \forall m \geq 1. \end{aligned} \tag{4}$$

In conjunction with the regular design $x_i = i/n$ for $1 \leq i \leq n$, this gives the collection of regressors

$$\phi_j = (\varphi_j(i/n))_{1\leq i\leq n}, \qquad 1 \leq j \leq k_n.$$

In practice, we suppose that f belongs to one of the periodic Sobolev classes:

DEFINITION 1. Let $\alpha > 0$ and $L > 0$. Let $(\varphi_j)_{j\geq 1}$ denote the Fourier basis (4). We define the Sobolev class $\mathcal{W}(\alpha, L)$ as the collection of all functions $f = \sum_{j=1}^{\infty} \theta_j \varphi_j$ in $\mathbb{L}^2([0,1])$ such that $\theta = (\theta_j)_{j\geq 1}$ is an element of the ellipsoid of $\ell^2(\mathbb{N})$,

$$\Theta(\alpha, L) = \left\{ \theta \in \ell^2(\mathbb{N}) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq \frac{L^2}{\pi^{2\alpha}} \right\},$$

where

$$a_j = \begin{cases} j^\alpha, & \text{if } j \text{ is even;} \\ (j-1)^\alpha, & \text{if } j \text{ is odd.} \end{cases} \tag{5}$$

(3) *Regression of a function in* $C^\alpha[0,1]$. Let $\alpha > 0$, and $f \in C^\alpha[0,1]$. This means that f is α_0 times continuously differentiable with $\|f\|_\alpha < \infty$, α_0 being the greatest integer less than α and the seminorm $\|\cdot\|_\alpha$ being defined by

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

Consider a design $(x_i^{(n)})_{n \geq 1, 1 \leq i \leq n}$, not necessarily uniform. Here F_0 is the vector $(f(x_i^{(n)}))_{1 \leq i \leq n}$. Once again we suppose that $\sigma_n = \sigma$ does not depend on n.

Fix an integer $q \geq \alpha$, and let $K = k_n + 1 - q$. Partition the interval $(0,1]$ into K subintervals $((j-1)/K, j/K]$ for $1 \leq j \leq K$. We want to perform the regression of f in the space of splines of order q defined on that partition, and use the B-splines basis $(B_j)_{1 \leq j \leq k_n}$ [see, e.g., de Boor (1978)]. Our collection of regressors is $\phi_j = (B_j(x_i^{(n)}))_{1 \leq i \leq n}$, for $1 \leq j \leq k_n$.

For any value of $n \geq 1$, let $\widetilde{W}$ be a prior distribution on $\mathbb{R}^{k_n}$ and, for $F = \Phi\theta$, let W be the prior distribution on $F \in \mathbb{R}^n$ obtained from $\widetilde{W}$ on θ. Its support is included in $\langle\phi\rangle$. Let P^W denote the marginal distribution of Y under prior W, and $W(dG(F)|Y)$ denote the posterior distribution of a functional $G(F)$. Note that everything depends on n (W, e.g., is a distribution on $\mathbb{R}^n$) even if we do not use n as an index to simplify our notation.

Both the parametrization by θ and the corresponding collection of regressors $\phi_1, \ldots, \phi_{k_n}$ are arbitrary: what matters is the posterior distribution of F and this depends on the space $\langle\phi\rangle$, not on the basis used to parametrize it. The span $\langle\phi\rangle$ is characterized by the matrix $\Sigma = \Phi(\Phi^T\Phi)^{-1}\Phi^T$ of the orthogonal projection onto $\langle\phi\rangle$.

The prior W is a sieve prior: that is, its support comes from a finite-dimensional model whose dimension k_n grows with n. The collection of growing models $\langle\phi\rangle$ (the sieve) can be seen as an approximation framework, each model being possibly misspecified. There is no true parameter in our setting: the true mean vector F_0 may fall outside $\langle\phi\rangle$ and correspond to none of the possible values of θ. There is then a bias which has to be dealt with, linked to the choice of the cutoff k_n.

When dealing with Bernstein–von Mises results, the question of the asymptotic centering point arises. In nonparametric models constructed on an infinite-dimensional parameter, there is no definition of a MLE; what the

natural centering for a Bernstein–von Mises theorem should be in such situations is not clear. In the model $\langle\phi\rangle$, the orthogonal projection $Y_{\langle\phi\rangle} = \Sigma Y$ of Y is also the MLE of F_0. We set $\theta_Y = (\Phi^T\Phi)^{-1}\Phi^T Y$ its associated parameter. Let also $F_{\langle\phi\rangle} = \Phi\theta_0$ be the projection of F_0 on $\langle\phi\rangle$, with $\theta_0 = (\Phi^T\Phi)^{-1}\Phi^T F_0$. Now, $F_0 - F_{\langle\phi\rangle}$ corresponds to the bias introduced by the use of the model $\langle\phi\rangle$, and $F_{\langle\phi\rangle}$ is the centering point of the distribution of the MLE $Y_{\langle\phi\rangle}$ under P_{F_0}:

$$Y_{\langle\phi\rangle} \sim \mathcal{N}(F_{\langle\phi\rangle}, \sigma_n^2\Sigma).$$

Although the MLE is naturally defined *in the sieve* $\langle\phi\rangle$, it heavily depends on the choice of $\langle\phi\rangle$. Therefore, the Bernstein–von Mises theorems we establish depend on the choice of the sieve the prior distribution is built on.

3. Nonparametric Bernstein–von Mises theorems. The proofs of our nonparametric results are delayed to Section 6.

3.1. *With Gaussian priors.* We consider here a centered, normal prior distribution W which is isotropic on $\langle\phi\rangle$, so that $W = \mathcal{N}(0, \tau_n^2\Sigma)$ for some sequence τ_n. τ_n is a scale parameter, and essentially the only assumption needed in this case is that τ_n is large enough as n grows. Let $\|Q - Q'\|_{\mathrm{TV}}$ denote the total variation norm between two probability distributions Q and Q'.

THEOREM 1. *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*

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

In terms of θ instead of F, an equivalent statement is

$$E\|\widetilde{W}(d\theta|Y) - \mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \to 0 \qquad \text{as } n \to \infty.$$

Theorem 1 does not deal with the modeling bias introduced by taking a prior restricted to $\langle\phi\rangle$. This is an important question in nonparametric statistics, and k_n has to be chosen in order to achieve a satisfactory bias-variance trade-off.

As an example, let us consider a typical regression framework with $F_0 = (f_0(x_i))_{1\le i\le n}$, where f_0 is some function and $(x_i)_{1\le i\le n}$ some design. If σ_n does not depend on n, both conditions $\|F_0\| = o(\tau_n^2/\sigma_n)$ and $k_n = o(\tau_n^4/\sigma_n^4)$ are satisfied if f_0 is bounded and $n^{1/4} = o(\tau_n)$. These conditions can be read in another way: τ_n^4 must be large enough with respect to $\|F_0\|$ and k_n.

3.2. *With smooth priors.* We consider now more general priors. To understand better the conditions we use, we need to look at the mechanics of the Bernstein–von Mises theorem.

Behind a Bernstein–von Mises theorem there is a LAN structure: the log-likelihood admits a quadratic expansion near the MLE. Since the posterior density is proportional to the product of the prior density and the likelihood, the prior has to be locally constant to let the likelihood alone influence the posterior and produce the Gaussian shape. To prove a Bernstein–von Mises theorem, we look for a subset which is simultaneously (1) large enough, so that the posterior will concentrate on it, and (2) small enough, so that we can find approximately constant priors on it. The larger the dimension of the model is, the more difficult it is to combine these two requirements, and the more difficult it is to obtain a Bernstein–von Mises theorem.

The geometry of the subsets are naturally suggested by the normal distribution we are looking for. For $M > 0$, consider the ellipsoid

$$\mathcal{E}_{\theta_0,\Phi}(M) = \{\theta \in \mathbb{R}^{k_n} : (\theta - \theta_0)^T \Phi^T \Phi (\theta - \theta_0) \le \sigma_n^2 M\}. \tag{6}$$

THEOREM 2. *Suppose that W is induced by a distribution $\widetilde{W}$ on θ admitting a density $w(\theta)$ with respect to the Lebesgue measure. If there exists a sequence $(M_n)_{n\ge 1}$ such that:*

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

then

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

With condition (1) below we ask for a sufficiently flat prior $\widetilde{W}$ in an ellipsoid $\mathcal{E}_{\theta_0,\Phi}(M_n)$. Condition (2) ensures, in particular, that the weight the normal distribution puts on $\mathcal{E}_{\theta_0,\Phi}(M_n)$ in the limit goes to 1. Condition (3) makes quantities linked to the volume of $\mathcal{E}_{\theta_0,\Phi}(M_n)$ appear and guarantees that it has enough prior weight. This kind of assumption is common in the literature dealing with the concentration of posterior distributions; see, for instance, Ghosal, Ghosh and van der Vaart (2000).

Several of our applications illustrate that priors known to induce the posterior minimax convergence rate may not be flat enough to get the Gaussian shape with the asymptotic variance $\sigma_n^2\Sigma$.

An important remark is the following: condition (2) does not really limit the growth rate of k_n. Read in conjunction with the other two conditions, we see that a flatter prior distribution will permit us to take M_n larger. Thus, the only condition on the growth rate of k_n is $k_n \le n$.

Note that Theorem 2 is not a generalization of Theorem 1: Theorem 1 is more powerful for isotropic Gaussian priors. Consider again the regression framework with $F_0 = (f_0(x_i))_{1\le i\le n}$, where f_0 is a bounded function and $(x_i)_{1\le i\le n}$ is some design. Suppose that σ_n does not depend on n, and take $k_n = n$ and $W = \mathcal{N}(0, \tau_n^2\Sigma)$. Then the conditions of Theorem 1 are satisfied as soon as $n^{1/4} = o(\tau_n)$, but with Theorem 2 we need $n \ln n = o(\tau_n^2)$.

Our main applications, to the Gaussian sequence model and to the regression model using trigonometric polynomials and splines, are developed in Section 5. We now present two remarks about the parametric case and the comparison with the pioneer work of Ghosal (1999).

The parametric case. Consider the regression of a function f defined on $[0,1]$, with a fixed number k of regressors. Set a design $(x_i^{(n)})_{n\ge 1, 1\le i\le n}$, with $x_i^{(n)} \in [(i-1)/n, i/n]$ for any $n \ge 1$, and $F_0 = (f(x_i^{(n)}))_{1\le i\le n}$. Choose a finite number of piecewise continuous and linearly independent regressors $(\varphi_j)_{1\le j\le k}$ on $[0,1]$, and set $\phi_j = (\varphi_j(x_i^{(n)}))_{1\le i\le n}$ for $1 \le j \le k$. Assume that f, $k_n = k$, $\sigma_n = \sigma$ and $\widetilde{W}$ do not depend on n.

We would like to compare Theorem 2 with the usual Bernstein–von Mises theorem for parametric models applied to such a regression framework. In that setting, let us suppose that w is continuous and positive, and that f is bounded. Then condition (1) becomes $M_n = o(n)$, while condition (3) reduces to $\ln n = o(M_n)$. Clearly, there exist such sequences $(M_n)_{n\ge 1}$, so Theorem 2 applies. Here the rescaling by $\sqrt{n}$ of the Bernstein–von Mises theorem for parametric models is hidden in the asymptotic posterior variance $\sigma^2(\Phi^T\Phi)^{-1}$ of the parameter θ. Indeed, $(1/n)\ \Phi^T\Phi$ is a Riemann sum and converges toward the Gramian matrix of the collection $(\varphi_j)_{1\le j\le k}$ in $\mathbb{L}^2([0,1])$.

Proof. We have $\|\Phi\theta_0\| \le \|F_0\| \le \sqrt{n}\|f\|_\infty$, and $\|\theta_0\|^2 \le \|(\Phi^T\Phi)^{-1}\| \cdot \|\Phi\theta_0\|^2 \le \|n(\Phi^T\Phi)^{-1}\|\|f\|_\infty^2$. $(1/n)\ \Phi^T\Phi$ converges toward the Gramian matrix of the collection $(\varphi_j)_{1\le j\le k}$ in $\mathbb{L}^2([0,1])$, and its smallest eigenvalue is lower bounded for n large enough. Therefore, θ_0 is bounded, and we can consider it lies in some compact set on which w is uniformly continuous and lower bounded by a positive constant. The rest follows. □

Comparison with Ghosal's conditions. The Bernstein–von Mises theorem in a regression setting when the number of parameters goes to infinity has been first studied by Ghosal (1999) as an early step in the development of frequentist nonparametric Bayesian theory. In his paper the errors ε_i are not supposed to be Gaussian. Under the Gaussianity assumption we get improved results, which means that we have a nontrivial generalization of the Ghosal (1999) conditions in the case of Gaussian errors. In particular,

our condition for the prior smoothness is simpler, and the growth rate of the dimension k_n is much less constrained:

- Ghosal (1999) does not admit a modeling bias between F_0 and $\Phi\theta_0$. In the present work the normality of the errors permits us to take $F_0 \neq \Phi\theta_0$ without any cost, as it appears in the core of the proof (Lemma 7). The possibility of considering misspecified models is an important improvement.
- In Ghosal (1999) σ_n is constant, which does not allow the application to the Gaussian sequence model.
- Ghosal (1999) restricts the growth of the dimension k_n to $k_n^4 \ln k_n = o(n)$ (see below). In our setting we only require $k_n \leq n$. With Ghosal's condition we could not have obtained the applications to the Gaussian sequence model or to the regression model for Sobolev or C^α classes.

Let $\delta_n^2 = \|(\Phi^T\Phi)^{-1}\|$ be the operator norm of $(\Phi^T\Phi)^{-1}$ for the ℓ^2 metric, and let η_n^2 be the maximal value on the diagonal of Σ. With our notation, the last two assumptions of Ghosal (1999) become:

(A3) There exists $\eta_0 > 0$ such that $w(\theta_0) > \eta_0^{k_n}$. Moreover,

$$|\ln w(\theta) - \ln w(\theta_0)| \leq L_n(C)\|\theta - \theta_0\|, \tag{7}$$

whenever $\|\theta - \theta_0\| \leq C\delta_n k_n \sqrt{\ln k_n}$, where the Lipschitz constant $L_n(C)$ is subject to some growth restriction [see assumption (A4)].

(A4)

$$\forall C > 0 \qquad L_n(C)\delta_n k_n \sqrt{\ln k_n} \to 0 \quad \text{and} \quad \eta_n k_n^{3/2} \sqrt{\ln k_n} \to 0. \tag{8}$$

Further, the design satisfies a condition on the trace of $\Phi^T\Phi$:

$$\operatorname{tr}(\Phi^T\Phi) = O(nk_n). \tag{9}$$

Since Σ is an orthogonal projection matrix on a k_n-dimensional space, $\operatorname{tr}(\Sigma) = k_n$ and $\eta_n^2 \geq k_n/n$. Thus, the last part of (8) implies $k_n^4 \ln k_n = o(n)$.

If we add the normality of the errors and a slight technical condition $\ln n = o(k_n \ln k_n)$, these assumptions imply ours. Indeed, set $M_n = C^2 k_n^2 \ln k_n$ for some arbitrary value of C. Our condition (2) is immediate. Condition (1) is got from (7) and the first part of (8). The beginning of (A3) implies $-\ln w(\theta_0) = O(k_n) = o(M_n)$. Using the concavity of the ln function and (9), we get $\ln\det(\Phi^T\Phi) \leq k_n \ln \operatorname{tr}(\Phi^T\Phi) - k_n \ln k_n = O(k_n \ln n) = o(M_n)$. Therefore, our condition (3) holds.

4. Semiparametric Bernstein–von Mises theorems. We consider two kinds of functionals of F: linear and nonlinear ones. These results can be easily adapted to functionals of θ, using the maps $\theta \mapsto \Phi\theta$ and $F \mapsto (\Phi^T\Phi)^{-1}\Phi^T F$.

4.1. *The linear case.* For linear functionals of F, we have the following corollary:

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

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

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

Corollary 1 is just a linear transform of the preceding theorems, and of the distribution of $Y_{\langle\phi\rangle}$.

An example of application is given in Section 5.2, in the context of the regression on Fourier's basis.

4.2. *The nonlinear case.* The Bernstein–von Mises theorem which is presented here for nonlinear functionals is derived from the nonparametric theorems thanks to Taylor expansions. In the Taylor expansion of a functional, the first order term naturally leads to the posterior normality, as in the case of linear functionals. We do not want that the second order term interfere with this phenomenon: it has to be controlled. The conditions of Theorem 2 below are stated to permit this control of the second order term.

Let $p \geq 1$ be fixed, and $G : \mathbb{R}^n \mapsto \mathbb{R}^p$ be a twice continuously differentiable function. For $F \in \mathbb{R}^n$, let $\dot{G}_F$ denote the Jacobian matrix of G at F, and $D_F^2 G(\cdot,\cdot)$ the second derivative of G, as a bilinear function on $\mathbb{R}^n$. For any $F \in \langle\phi\rangle$ and $a > 0$, let

$$B_F(a) = \sup_{h \in \langle\phi\rangle\,:\,\|h\|^2 \leq \sigma_n^2 a} \sup_{0 \leq t \leq 1} \|D^2_{F+th} G(h,h)\|, \tag{10}$$

where $\|\cdot\|$ denotes the Euclidean norm of $\mathbb{R}^p$.

We also consider the following nonnegative symmetric matrix

$$\Gamma_F = \sigma_n^2 \dot{G}_F \Sigma \dot{G}_F^T. \tag{11}$$

In the following, $\|\Gamma_F^{-1}\|$ denotes the Euclidean operator norm of Γ_F^{-1}, which is also the inverse of the smallest eigenvalue of Γ_F.

Let $\mathcal{I}$ be the collection of all intervals in $\mathbb{R}$, and for any $I \in \mathcal{I}$, let $\psi(I) = P(Z \in I)$, where Z is a $\mathcal{N}(0,1)$ random variable. Recall that $Y_{\langle\phi\rangle}$ is the MLE and the orthogonal projection of Y on $\langle\phi\rangle$.

THEOREM 3. *Let $G : \mathbb{R}^n \mapsto \mathbb{R}^p$ be a twice continuously differentiable function, and let Γ_F be as just defined. 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*

$$B^2_{F_{\langle\phi\rangle}}(M_n) = o(\|\Gamma^{-1}_{F_{\langle\phi\rangle}}\|^{-1}). \tag{12}$$

Suppose further that the conditions of either Theorems 1 or 2 are satisfied. Then, for any $b \in \mathbb{R}^p$,

$$E\left[\sup_{I\in\mathcal{I}}\left|W\left(\frac{b^T(G(F)-G(Y_{\langle\phi\rangle}))}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}}\in I\middle|Y\right)-\psi(I)\right|\right]\to 0 \qquad \text{as } n\to\infty. \tag{13}$$

Under the same conditions,

$$\sup_{I\in\mathcal{I}}\left|P\left(\frac{b^T(G(Y_{\langle\phi\rangle})-G(F_{\langle\phi\rangle}))}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}}\in I\right)-\psi(I)\right|\to 0 \qquad \text{as } n\to\infty. \tag{14}$$

Note that $\sup_{I\in\mathcal{I}}|Q(I)-Q'(I)|$ is the Levy–Prokhorov distance between two distributions Q and Q' on $\mathbb{R}$. The Levy–Prokhorov distance metrizes the convergence in distribution. So, when $p=1$ the Levy–Prokhorov distance between the distribution $W(dG(F)|Y)$ and $\mathcal{N}(G(Y_{\langle\phi\rangle}),\Gamma_{F_{\langle\phi\rangle}})$ goes to 0 in mean, while $G(Y_{\langle\phi\rangle})$ goes to $\mathcal{N}(G(F_{\langle\phi\rangle}),\Gamma_{F_{\langle\phi\rangle}})$ in distribution.

An application of Theorem 3 is given in Section 5.2, in the context of the regression on Fourier's basis. The proof is delayed to Section 6.3.

5. Applications. Here we give the three applications described in Section 2. The models studied and the collections of regressors used have already been defined there.

5.1. *The Gaussian sequence model.* We consider the model (2). Here the MLE is the projection $\theta_Y=(Y_j)_{1\le j\le k_n}$.

The nonparametric case corresponds to the estimation of θ^0. Under the assumption that θ^0 is in some regularity class, we will obtain a Bernstein–von Mises theorem with the posterior convergence rate already obtained in previous works, in particular, Ghosal and van der Vaart (2007). On the other hand, for some priors known to achieve this rate, it will be seen that the centering point and the asymptotic variance of the posterior distribution do not fit with the ones expected in a Bernstein–von Mises theorem. We also look at the semiparametric estimation of the squared ℓ^2 norm of θ^0.

5.1.1. *The nonparametric estimation of θ^0.*

PROPOSITION 1. *Suppose that $\sum_{j=1}^{k_n}(\theta_j^0)^2$ is bounded. This holds when θ^0 is an element of $\ell^2(\mathbb{N})$ not depending on n. With a prior $\widetilde{W}=\mathcal{N}(0,\tau_n^2 I_{k_n})$ such that $n^{-1/4}=o(\tau_n)$, we have for any sequence $k_n\le n$,*

$$E\left\|\widetilde{W}(d\theta|Y)-\mathcal{N}\left(\theta_Y,\frac{1}{n}I_{k_n}\right)\right\|_{\mathrm{TV}}\to 0 \qquad \text{as } n\to\infty,$$

and the convergence rate of θ *toward* θ_0 *is* $\sqrt{\frac{k_n}{n}}$*: for every* $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|\theta-\theta_0\| \geq \lambda_n\sqrt{\frac{k_n}{n}}\Big|Y\right)\right] \to 0.$$

Recall that $\theta_0 = (\theta_j^0)_{1\leq j\leq k_n}$ is the projection of θ^0.

Proof of Proposition 1. The beginning is an immediate corollary of Theorem 1. For the convergence rate, let $\lambda_n \to \infty$. Since $\theta_Y - \theta_0 \sim \mathcal{N}(0, \frac{1}{n}I_{k_n})$,

$$P\left(\|\theta_Y-\theta_0\| \geq \frac{\lambda_n}{2}\sqrt{\frac{k_n}{n}}\right) \to 0.$$

In the same way

$$E\left[\widetilde{W}\left(\|\theta-\theta_Y\| \geq \frac{\lambda_n}{2}\sqrt{\frac{k_n}{n}}\right)\right] \leq E\left\|\widetilde{W}(d\theta|Y) - \mathcal{N}\left(\theta_Y, \frac{1}{n}I_{k_n}\right)\right\|_{\mathrm{TV}} + \mathcal{N}\left(0, \frac{1}{n}I_{k_n}\right)\left(\left\{\|h\| \leq \frac{\lambda_n}{2}\sqrt{\frac{k_n}{n}}\right\}\right),$$

which goes to 0. Therefore,

$$E\left[\widetilde{W}\left(\|\theta-\theta_0\| \geq \lambda_n\sqrt{\frac{k_n}{n}}\right)\right] \to 0. \qquad \square$$

However, in such a general setting we have no information about the bias between θ^0 and its projection θ_0. Several authors add the assumption that the true parameter belongs to a Sobolev class of regularity $\alpha > 0$, defined by the relation $\sum_{j=1}^{\infty}|\theta_j^0|^2 j^{2\alpha} < \infty$. In this setting we show that for some priors the induced posterior may achieve the nonparametric convergence rate but with a centering point and a variance different from what is expected in the Bernstein–von Mises theorem. Then we exhibit priors for which both the Bernstein–von Mises theorem and the nonparametric convergence rate hold.

From now on, we suppose that $\sum_{j=1}^{\infty}|\theta_j^0|^2 j^{2\alpha} < \infty$. In this setting Ghosal and van der Vaart (2007), Section 7.6, consider a prior $\widetilde{W}$ such that θ_1, $\theta_2, \ldots$ are independent, and θ_j is normally distributed with variance σ_{j,k_n}^2. Further, the variances are supposed to satisfy

$$c/k_n \leq \min\{\sigma_{j,k_n}^2 j^{2\alpha} : 1 \leq j \leq k_n\} \leq C/k_n \tag{15}$$

for some positive constants c and C. Suppose that $\alpha \geq 1/2$ and there exist constants C_1 and C_2 such that $C_1 n^{1/(1+2\alpha)} \leq k_n \leq C_2 n^{1/(1+2\alpha)}$. Then Ghosal

and van der Vaart (2007), Theorem 11, proved that the posterior converges at the rate $n^{-\alpha/(1+2\alpha)}$.

In order to get $n^{-1}I_{k_n}$ as asymptotic variance, we need more stringent conditions on k_n, or a flatter prior. To see this is necessary, consider, for $k_n \approx n^{1/(1+2\alpha)}$, the following choice for σ_{j,k_n}:

$$\sigma_{j,k_n}^2 = \begin{cases} k_n^{-1}, & \text{if } 1 \le j \le k_n/2, \\ 2^{2\alpha}/n, & \text{if } j > k_n/2. \end{cases}$$

Then $\min\{\sigma_{j,k_n}^2 j^{2\alpha} : 1 \le j \le k_n\} \approx k_n^{-1}$, and the posterior converges at the rate $n^{-\alpha/(1+2\alpha)}$.

For this case we can explicitly calculate the posterior distribution. This is similar to the calculation made in the proof of Theorem 1. The coordinates are independent, and

$$\widetilde{W}(d\theta_j|Y) = \mathcal{N}\left(\frac{\sigma_{j,k_n}^2}{\sigma_n^2+\sigma_{j,k_n}^2}Y_j, \frac{\sigma_n^2\sigma_{j,k_n}^2}{\sigma_n^2+\sigma_{j,k_n}^2}\right).$$

For $j > k_n/2$, $\frac{\sigma_{j,k_n}^2}{\sigma_n^2+\sigma_{j,k_n}^2} = \frac{4^\alpha}{1+4^\alpha}$, and, therefore, $\|\widetilde{W}(d\theta_j|Y) - \mathcal{N}(Y_j, \sigma_n^2)\|_{\mathrm{TV}}$ is bounded away from 0.

By contrast, with an isotropic and flat prior we obtain the centering point and the asymptotic variance we expected, and the same convergence rate as previously. We have the following:

PROPOSITION 2. *Suppose that θ^0 belongs to the Sobolev class of regularity $\alpha > 0$. Choose a prior $\widetilde{W} = \mathcal{N}(0, \tau_n^2 I_{k_n})$ such that $n^{-1/4} = o(\tau_n)$, which ensures the asymptotic normality of the posterior distribution as in Proposition 1.*

If further $k_n \approx n^{1/(1+2\alpha)}$, then the convergence rate of θ toward θ_0 and toward θ^0 is $n^{-\alpha/(1+2\alpha)}$: for every $\lambda_n \to \infty$,

$$E[\widetilde{W}(\|\theta - \theta^0\| \ge \lambda_n n^{-\alpha/(1+2\alpha)}|Y)] \to 0.$$

PROOF. We consider θ and θ_0 as elements of $\ell^2(\mathbb{N})$ by setting $\theta_j = \theta_{0,j} = 0$ for $j \ge k_n + 1$. The convergence rate toward θ_0 has already been established in Proposition 1. Since $\theta_{0,j} = \theta_j^0$ for $1 \le j \le k_n$, $\|\theta^0 - \theta_0\| \le k_n^{-\alpha}\sqrt{\sum_{j=k_n+1}^{\infty}(\theta_j^0)^2 j^{2\alpha}} = O(k_n^{-\alpha})$. Therefore, the convergence rate of θ toward θ^0 is also $n^{-\alpha/(1+2\alpha)}$. □

5.1.2. *Semiparametric theorem for the ℓ^2 norm of θ^0.* We consider the prior distribution used in Proposition 2, but now we look at the posterior distribution of $\|\theta\|^2$. To get asymptotic normality with variance $n^{-1/2}$, we just need $k_n = o(\sqrt{n})$. To control the bias term, we need $\alpha > 1/2$, and in this case we get an adaptive Bayesian estimator.

PROPOSITION 3. *Let $\alpha > 1/2$ and suppose that θ^0 belongs to the Sobolev class of regularity α. Choose a prior $\widetilde{W} = \mathcal{N}(0, \tau_n^2 I_{k_n})$ such that $n^{-1/4} = o(\tau_n)$. Then, for any choice of k_n such that $k_n = o(\sqrt{n})$ and $\sqrt{n} = o(k_n^{2\alpha})$,*

$$E\left[\sup_{I\in\mathcal{I}}\left|\widetilde{W}\left(\frac{\sqrt{n}(\|\theta\|^2 - \|\theta_Y\|^2)}{2\|\theta^0\|} \in I \middle| Y\right) - \psi(I)\right|\right] \to 0 \qquad \text{as } n \to \infty$$

and $\frac{\sqrt{n}(\|\theta_Y\|^2 - \|\theta_0\|^2)}{2\|\theta^0\|} \to \mathcal{N}(0,1)$ in distribution, as $n \to \infty$. Further, the bias is negligible with respect to the square root of the variance:

$$\frac{\sqrt{n}(\|\theta_0\|^2 - \|\theta^0\|^2)}{2\|\theta^0\|} = o(1).$$

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

PROOF. We set up an application of Theorem 3. Since $\sigma_n = n^{-1/2}$, the conditions of Theorem 1 are fulfilled.

Here $G(\theta) = \theta^T\theta$, $\dot{G}_\theta = 2\theta^T$ and $\ddot{G}_\theta = 2I_{k_n}$. Therefore, $B_{\theta_0}(M_n) = 2M_n/n$, while $\Gamma_{\theta_0} = 4\|\theta_0\|^2/n$.

Let us choose $(M_n)_{n\geq 1}$ such that $k_n = o(M_n)$ and $M_n = o(\sqrt{n})$. Such sequences exist and fulfill the conditions of Theorem 3.

Since $\|\theta_0\|^2 \to \|\theta^0\|^2$, we can substitute the variance Γ_{θ_0} by $4\|\theta^0\|^2/n$ and get the two asymptotic normality results, (13) and (14).

As $n \to \infty$, $\|\theta^0\|^2 - \|\theta_0\|^2 = \|\theta^0 - \theta_0\|^2 = O(k_n^{-2\alpha})$, as in the proof of Proposition 2. If $\sqrt{n} = o(k_n^{2\alpha})$, we get $\sqrt{n}(\|\theta_0\|^2 - \|\theta^0\|^2) = o(1)$. □

5.2. *Regression on Fourier's basis.* Now we consider the regression model (3) with a function f in a Sobolev class $\mathcal{W}(\alpha, L)$, and use Fourier's basis (4). For any $\theta \in \mathbb{R}^{k_n}$, we define $f_\theta = \sum_{j=1}^{k_n} \theta_j \varphi_j$. We also denote by $\theta^0 \in \ell^2(\mathbb{N})$ the sequence of Fourier's coefficients of f: $f = \sum_{j=1}^{\infty} \theta_j^0 \varphi_j$.

The following useful lemma about our collection of regressors can be found, for instance, in Tsybakov (2004) (we slightly modified it to take into account the case n even):

LEMMA 1. *Suppose either that n is odd and $k_n \leq n$, or n is even and $k_n \leq n-1$. Consider the collection $(\phi_j)_{1\leq j\leq k_n}$ defined before, and Φ the associated matrix. Then*

$$\Phi^T\Phi = nI_{k_n}.$$

This makes the regression on Fourier's basis very close to the Gaussian sequence model, and the results we obtain are similar.

In this subsection we first consider the estimation of f in a Sobolev class, for which we get a Bernstein–von Mises theorem and the frequentist minimax

$n^{-\alpha/(1+2\alpha)}$ posterior convergence rate for the L^2 norm. Then we consider two semiparametric settings: the estimation of a linear functional of f, and the estimation of the L^2 norm of f. We get the adaptive $\sqrt{n}$ convergence rate for any $\alpha > 1/2$.

5.2.1. *Nonparametric Bernstein–von Mises theorem in Sobolev classes.*

PROPOSITION 4. *Suppose that f belongs to some Sobolev class $\mathcal{W}(\alpha, L)$ for $L > 0$ and $\alpha > 1/2$. Let $k_n \approx n^{1/(1+2\alpha)}$ and $\widetilde{W} = \mathcal{N}(0, \gamma_n I_{k_n})$ be the prior on θ, for a sequence $(\gamma_n)_{n\geq 1}$ such that $1/\sqrt{n} = o(\gamma_n)$. Then*

$$E\left\|\widetilde{W}(d\theta|Y) - \mathcal{N}\left(\theta_Y, \frac{\sigma^2}{n} I_{k_n}\right)\right\|_{\mathrm{TV}} \to 0 \qquad \text{as } n \to \infty$$

and the convergence rate relative to the Euclidean norm for f_θ is $n^{-\alpha/(1+2\alpha)}$: for every $\lambda_n \to \infty$,

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

PROOF. The conditions of Theorem 1 are fulfilled: with $\tau_n^2 = n\gamma_n$, we have $n = o(\tau_n^4)$. The first assertion follows.

Because of the orthogonal nature of Fourier's basis, $\|f_\theta - f\| = \|\theta - \theta^0\|$ in $\ell^2(\mathbb{N})$. We use the decomposition $\|\theta - \theta^0\|^2 \leq \|\theta - \theta_0\|^2 + \|\theta_0 - \theta^0\|^2$. In the same way as in the proof of Proposition 1, for any $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|\theta - \theta_0\| \geq \lambda_n \sqrt{\frac{k_n}{n}}\right)\right] \to 0.$$

Going back to Definition 1, we have

$$\|\theta_0 - \theta^0\|^2 = \sum_{j=k_n+1}^{\infty} (\theta_j^0)^2 \leq k_n^{-2\alpha} \sum_{j=k_n+1}^{\infty} a_j^{2\alpha} (\theta_j^0)^2 = O(k_n^{-2\alpha}).$$

This permits to get

$$E[\widetilde{W}(\|\theta - \theta^0\| \geq \lambda_n n^{-\alpha/(1+2\alpha)}|Y)] \to 0. \qquad \square$$

5.2.2. *Linear functionals of f.* Let $g : [0,1] \to \mathbb{R}$ be a function in $\mathbb{L}^2([0,1])$. We want to estimate $\mathcal{F}(f) = \int_0^1 fg$, and we approximate it by

$$\frac{1}{n} \sum_{i=1}^{n} g(i/n) f(i/n) = GF_0,$$

where $G = (g(i/n)/n)_{1\leq i\leq n}^T$. The plug-in MLE estimator of GF_0 in the misspecified model $\langle\phi\rangle$ is $GY_{\langle\phi\rangle}$. More generally, we consider the functional $F \mapsto GF$. The following result is adaptive, in the sense that the same choice $k_n = \lfloor n/\ln n \rfloor$ entails the convergence rate $n^{-1/2}$ for all values of $\alpha > 1/2$.

PROPOSITION 5. *Suppose f is bounded, and let W be the prior induced by the $\mathcal{N}(0,\gamma_n\ I_{k_n})$ distribution on θ, for a sequence $(\gamma_n)_{n\geq 1}$ such that $1/\sqrt{n}=o(\gamma_n)$. Then:*

(1)

$$E\|W(d(GF)|Y)-\mathcal{N}(GY_{\langle\phi\rangle},\sigma^2G\Sigma G^T)\|_{\mathrm{TV}}\to 0$$

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

(2) *Suppose further that f and g belong to some Sobolev class $\mathcal{W}(\alpha,L)$ for $L>0$ and $\alpha>1/2$. Then $G\Sigma G^T\sim\frac{1}{n}\int_0^1 g^2$,*

$$E\left\|W\left(d\frac{\sqrt{n}(GF-GY_{\langle\phi\rangle})}{\sigma\sqrt{\int_0^1 g^2}}\Big|Y\right)-\mathcal{N}(0,1)\right\|_{\mathrm{TV}}\to 0$$

and $\frac{\sqrt{n}(GY_{\langle\phi\rangle}-GF_{\langle\phi\rangle})}{\sigma\sqrt{\int_0^1 g^2}}\to\mathcal{N}(0,1)$ in distribution, as $n\to\infty$.

(3) *Suppose that f and g belong to some Sobolev class $\mathcal{W}(\alpha,L)$ for $L>0$ and $\alpha>1/2$, and suppose further that k_n is large enough so that $n=o(k_n^{2\alpha})$. Then the bias is negligible with respect to the square root of the variance:*

$$\frac{\sqrt{n}(GF_{\langle\phi\rangle}-\mathcal{F}(f))}{\sigma\sqrt{\int_0^1 g^2}}=o(1).$$

Before the proof we give two lemmas, proved in Appendix B in the supplemental article [Bontemps (2011)], about the error terms of the approximation of a Sobolev class by a sieve build on Fourier's basis, and of the approximation of an integral by a Riemann sum.

LEMMA 2. *Let $\alpha>1/2$ and $L>0$. We suppose n odd or $k_n<n$. If $f\in\mathcal{W}(\alpha,L)$,*

$$\|F_0-F_{\langle\phi\rangle}\|\leq(1+o(1))\frac{\sqrt{2}L}{\pi^\alpha}\frac{\sqrt{n}}{k_n^\alpha}.$$

Further, $\|F_0\|\sim\sqrt{n\int_0^1 f^2}$ and $\|F_0-F_{\langle\phi\rangle}\|=O(k_n^{-\alpha}\|F_0\|)$.

LEMMA 3. *Let two functions $f\in\mathcal{W}(\alpha,L)$ and $g\in\mathcal{W}(\alpha',L')$ for some $\alpha,\alpha'>1/2$ and two positive numbers L and L'. Then*

$$\left|\frac{1}{n}\sum_{i=1}^n f(i/n)g(i/n)-\int_0^1 fg\right|=O(n^{-\inf(\alpha,\alpha')}).$$

PROOF OF PROPOSITION 5. (1) The first assertion is just Corollary 1. The conditions of Theorem 1 are fulfilled, as in the proof of Proposition 4.

(2) If $g \in \mathcal{W}(\alpha, L)$ for $L > 0$ and $\alpha > 1/2$, $G\Sigma G^T = \|\Sigma G^T\|^2 \sim \|G^T\|^2$ by Lemma 2. In the meantime $\|G^T\|^2 = \frac{1}{n^2}\sum_{i=1}^n g^2(x_i) \sim \frac{1}{n}\int_0^1 g^2$ by Lemma 3. So $G\Sigma G^T \sim \frac{1}{n}\int_0^1 g^2$, and the variance in the formulas of Corollary 1 can be substituted with $\frac{1}{n}\int_0^1 g^2$.

(3) We decompose the bias into two terms, $|GF_0 - \mathcal{F}(f)|$ and $|GF_{\langle\phi\rangle} - GF_0|$, and show that both are $o(n^{-1/2})$. The first term is controlled by Lemma 3. For the last one, $|GF_{\langle\phi\rangle} - GF_0| \le \|G^T\|\|F_{\langle\phi\rangle} - F_0\|$. But $\|G^T\| = O(n^{-1/2})$, $\|F_{\langle\phi\rangle} - F_0\| = O(k_n^{-\alpha}\|F_0\|)$ by Lemma 2 and $\|F_0\| = O(\sqrt{n})$. We conclude thanks to the assumption $n = o(k_n^{2\alpha})$. □

5.2.3. *L^2 norm of f.* Suppose that we want to estimate $\mathcal{F}(f) = \int_0^1 f^2$. We can consider the plug-in MLE estimator

$$G(Y_{\langle\phi\rangle}) = \frac{1}{n}\|Y_{\langle\phi\rangle}\|^2 = \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^{k_n} \theta_{Y,j}\varphi_j(i/n)\right)^2.$$

More generally, we define, for any $F \in \mathbb{R}^n$,

$$G(F) = \frac{1}{n}\|F\|^2. \tag{16}$$

With a Gaussian prior, we obtain the following result, which is also adaptive: the same $k_n = \lfloor \sqrt{n}/\ln n \rfloor$ is suitable whatever $\alpha > 1/2$.

PROPOSITION 6. *Let $G(F) = \|F\|^2/n$. Suppose that $f \in \mathcal{W}(\alpha, L)$ for some $L > 0$ and $\alpha > 1/2$. Let W be the prior induced by the $\mathcal{N}(0, \gamma_n\, I_{k_n})$ distribution on θ, for a sequence $(\gamma_n)_{n\ge1}$ such that $1/\sqrt{n} = o(\gamma_n)$. The sequence $(k_n)_{n\ge1}$ can be chosen such that $k_n = o(\sqrt{n})$ and $\sqrt{n} = o(k_n^{2\alpha})$, and with such a choice,*

$$E\left[\sup_{I\in\mathcal{I}}\left|W\left(\frac{\sqrt{n}(G(F) - G(Y_{\langle\phi\rangle}))}{2\sigma\sqrt{\mathcal{F}(f)}} \in I \middle| Y\right) - \psi(I)\right|\right] \to 0 \qquad \text{as } n \to \infty$$

and $\frac{\sqrt{n}(G(Y_{\langle\phi\rangle}) - G(F_{\langle\phi\rangle}))}{2\sigma\sqrt{\mathcal{F}(f)}} \to \mathcal{N}(0,1)$ in distribution, as $n \to \infty$. Further, the bias is negligible with respect to the square root of the variance:

$$\frac{\sqrt{n}(G(F_{\langle\phi\rangle}) - \mathcal{F}(f))}{2\sigma\sqrt{\mathcal{F}(f)}} = o(1).$$

A similar corollary could be stated for a non-Gaussian prior.

PROOF OF PROPOSITION 6. First, let us note that the conditions of Theorem 1 are fulfilled, as in the proof of Proposition 4. Lemma 10 in Appendix B insures that f is bounded.

In this setting $\dot{G}_F = (2/n)F^T$ and $D_F^2 G(h,h) = (2/n)\|h\|^2$ for any $F \in \mathbb{R}^n$ and any $h \in \mathbb{R}^n$. Therefore, $B_F(a) = 2\sigma^2 a/n$, and $\Gamma_F = 4(\sigma^2/n^2)\|F\|^2$. By Lemma 2, $\|F_{\langle\phi\rangle}\|^2 \sim \|F_0\|^2 \sim n\mathcal{F}(f)$. Thus, $\Gamma_{F_{\langle\phi\rangle}} = 4(1+o(1))\mathcal{F}(f)/n$.

Let us choose $(M_n)_{n\geq 1}$ such that $k_n = o(M_n)$ and $M_n = o(\sqrt{n})$. Such sequences exist and fulfill the conditions of Theorem 3. We can substitute the variance $\Gamma_{F_{\langle\phi\rangle}}$ by $4\mathcal{F}(f)/n$ and get the two asymptotic normality results.

Let us now consider the bias term:

$$\mathcal{F}(f) - G(F_{\langle\phi\rangle}) \leq \frac{\|F_0\|^2 - \|F_{\langle\phi\rangle}\|^2}{n} + \left(\int_0^1 f^2 - \frac{1}{n}\sum_{i=1}^n f^2(i/n)\right).$$

We use Lemma 2 to control $\|F_0\|^2 - \|F_{\langle\phi\rangle}\|^2$, and Lemma 3 for the other term:

$$|\mathcal{F}(f) - G(F_{\langle\phi\rangle})| = O(k_n^{-2\alpha}) + O(n^{-\alpha}).$$

This is a $o(1/\sqrt{n})$ under the assumptions of Corollary 6. $\square$

5.3. *Regression on splines.* Here we consider the regression model for functions in $C^\alpha[0,1]$ with $\alpha > 0$, using splines, set up in Section 2. We first develop further the framework and the assumptions used here, and recall the previous result of Ghosal and van der Vaart (2007), Section 7.7.1, which obtains the posterior concentration at the frequentist minimax rate. Then we present two Bernstein–von Mises theorems: the first one with the same prior as Ghosal and van der Vaart (2007) but a stronger condition on k_n (or equivalently on α); the second one with a flatter prior, for which we obtain the minimax convergence rate in addition to the asymptotic Gaussianity of the posterior distribution.

To see this, we begin with some preliminaries. For any $\theta \in \mathbb{R}^{k_n}$, define $f_\theta = \sum_{j=1}^{k_n} \theta_j B_j$. The B-splines basis has the following approximation property: for any $\alpha > 0$, there exist $C_\alpha > 0$ such that, if $f \in C^\alpha[0,1]$, there exists $\theta^\infty \in \mathbb{R}^{k_n}$ satisfying

$$\|f - f_{\theta^\infty}\|_\infty \leq C_\alpha k_n^{-\alpha}\|f\|_\alpha. \tag{17}$$

We need the design $(x_i^{(n)})_{n\geq 1, 1\leq i\leq n}$ to be sufficiently regular and, as stressed in Ghosal and van der Vaart (2007), the spatial separation property of B-splines permits us to express the precise condition in terms of the covariance matrix $\Phi^T\Phi$. We suppose that there exist positive constants C_1 and C_2 such that, as n increases, for any $\theta \in \mathbb{R}^{k_n}$,

$$C_1\frac{n}{k_n}\|\theta\|^2 \leq \theta^T\Phi^T\Phi\theta \leq C_2\frac{n}{k_n}\|\theta\|^2. \tag{18}$$

Let us associate the norm $\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n |f(x_i)|^2}$ to the design. Note that $\sqrt{n}\|f_\theta\|_n = \|\Phi\theta\|$ if $\theta \in \mathbb{R}^{k_n}$. Under (18) we have a relation between

$\|\cdot\|_n$ and the Euclidean norm on the parameter space: for every θ_1 and θ_2,

$$C_1\|\theta_1-\theta_2\| \le \sqrt{k_n}\|f_{\theta_1}-f_{\theta_2}\|_n \le C_2\|\theta_1-\theta_2\|.$$

With these conditions Ghosal and van der Vaart (2007), Theorem 12, get the posterior concentration at the minimax rate. Take $\alpha \ge 1/2$, let $\widetilde{W} = \mathcal{N}(0, I_{k_n})$ be the prior on the spline coefficients, and suppose there exist constants C_3 and C_4 such that $C_3 n^{1/(1+2\alpha)} \le k_n \le C_4 n^{1/(1+2\alpha)}$. Then the posterior concentrates at the minimax rate $n^{-\alpha/(1+2\alpha)}$ relative to $\|\cdot\|_n$: for every $\lambda_n \to \infty$,

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

This is equivalent to a convergence rate $n^{(1-2\alpha)/(2(1+2\alpha))}$ relative to the Euclidean norm for θ:

$$E[\widetilde{W}(\|\theta-\theta_0\| \ge \lambda_n n^{(1-2\alpha)/(2(1+2\alpha))}|Y)] \to 0.$$

Indeed, (17) and the projection property imply

$$\|f_{\theta_0}-f\|_n \le \|f_{\theta^\infty}-f\|_n \le \|f_{\theta^\infty}-f\|_\infty \le C_\alpha \|f\|_\alpha k_n^{-\alpha}.$$

Now, with modified assumptions we get the Bernstein–von Mises theorem in two different settings. First, with the same prior as Ghosal and van der Vaart (2007):

PROPOSITION 7. *Assume that f is bounded, $k_n = o((\frac{n}{\ln n})^{1/3})$ and (18) holds. Let $\widetilde{W} = \mathcal{N}(0, I_{k_n})$ be the prior on the spline coefficients. Then*

$$E\|\widetilde{W}(d\theta|Y) - \mathcal{N}(\theta_Y, \sigma^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \to 0 \qquad \text{as } n\to\infty \tag{19}$$

and the convergence rate relative to the Euclidean norm for θ is $\frac{k_n}{\sqrt{n}}$.

Remarks. We need $\alpha > 1$ to get the Gaussian shape with the same convergence rate as in Ghosal and van der Vaart (2007). The conditions of Proposition 7 are satisfied, in particular, if there exist constants C_3 and C_4 such that $C_3 n^{1/(1+2\alpha)} \le k_n \le C_4 n^{1/(1+2\alpha)}$. In this case the convergence rate for θ is $n^{(1-2\alpha)/(2(1+2\alpha))}$.

PROOF OF PROPOSITION 7. We set up an application of Theorem 2. We can choose M_n such that $k_n \ln n = o(M_n)$ and $M_n = o(\frac{n}{k_n^2})$. Assumption (2) is then trivially satisfied.

From (18) we get $\|\Phi^T\Phi\| \le C_2\frac{n}{k_n}$ and $\|(\Phi^T\Phi)^{-1}\| \le C_1^{-1}\frac{k_n}{n}$. We have also $\ln\det(\Phi^T\Phi) \le k_n \ln C_2 + k_n \ln(\frac{n}{k_n}) = O(k_n \ln n) = o(M_n)$. Since $\theta_0 = \Phi(\Phi^T\Phi)^{-1}F_0$,

$$\|\theta_0\|^2 \le \frac{k_n}{C_1 n}\|F_0\|^2 \le \frac{\|f\|_\infty}{C_1}k_n.$$

Therefore, $-\ln w(\theta_0) = O(1) + \frac{1}{2}\|\theta_0\|^2 = O(k_n) = o(M_n)$, and assumption (3) holds.

Let $h \in \mathbb{R}^{k_n}$ such that $\|\Phi h\|^2 \le \sigma^2 M_n$. We have $\|h\|^2 \le \|(\Phi^T\Phi)^{-1}\| \cdot \|\Phi h\|^2 \le \frac{\sigma^2 k_n M_n}{C_1 n} = o(k_n^{-1})$. Therefore,

$$\sup_{\|\Phi h\|^2 \le \sigma^2 M_n} \left| \ln \frac{w(\theta_0 + h)}{w(\theta_0)} \right| \le \sup_{\|\Phi h\|^2 \le \sigma^2 M_n} \frac{\|h\|^2 + 2\|h\|\|\theta_0\|}{2} = o(1) \tag{20}$$

and assumption (1) follows.

Let us now prove the convergence rate. Let $\lambda_n \to \infty$. Then

$$P\left(\|\theta_Y - \theta_0\| \ge \frac{\lambda_n k_n}{2\sqrt{n}} \right) \le P\left(\|\Phi(\theta_Y - \theta_0)\|^2 \ge \frac{C_1 \lambda_n^2 k_n}{4} \right) \to 0$$

since $\|\Phi(\theta_Y - \theta_0)\|^2 \sim \sigma^2 \chi^2(k_n)$. In the same way

$$\begin{aligned} E\left[\widetilde{W}\left(\|\theta - \theta_Y\| \ge \frac{\lambda_n k_n}{2\sqrt{n}} \right) \right] &\le E\|\widetilde{W}(d\theta|Y) - \mathcal{N}(\theta_Y, \sigma^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \\ &\quad + \mathcal{N}(0, \sigma^2(\Phi^T\Phi)^{-1})\left(\left\{ h : \|h\| \le \frac{\lambda_n k_n}{2\sqrt{n}} \right\} \right) \\ &\to 0, \end{aligned}$$

where Theorem 2 controls the first term in the right. Therefore, assumption (3) holds:

$$E\left[\widetilde{W}\left(\|\theta - \theta_0\| \ge \frac{\lambda_n k_n}{\sqrt{n}} \right) \right] \to 0.$$

Now, (19) is the same as Theorem 2 in terms of $\widetilde{W}$. □

The situation is similar to the one we encountered with the Gaussian sequence model. To get the Bernstein–von Mises theorem with the same convergence rate as Ghosal and van der Vaart (2007) for $\alpha \le 1$, we need a flatter prior:

PROPOSITION 8. *Assume that f is bounded and (18) holds. Let $\widetilde{W} = \mathcal{N}(0, \tau_n^2 I_{k_n})$ be the prior on the spline coefficients, with the sequence τ_n satisfying*

$$\frac{k_n^2 \ln n}{n} = o(\tau_n^2) \quad \text{and} \quad \frac{k_n^3 \ln n}{n} = o(\tau_n^4).$$

Then

$$E\|\widetilde{W}(d\theta|Y) - \mathcal{N}(\theta_Y, \sigma^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \to 0 \qquad \text{as } n \to \infty$$

and the convergence rate relative to the Euclidean norm for θ is $\frac{k_n}{\sqrt{n}}$.

When $\alpha > 0$ and k_n is of order $n^{1/(1+2\alpha)}$, the conditions reduce to $n^{(2-2\alpha)/(1+2\alpha)} \ln n = o(\tau_n^4)$. So we obtain the convergence rate of Ghosal and

van der Vaart (2007) in addition to the Gaussian shape with the same k_n, even for $\alpha \le 1$, but with a different prior.

PROOF OF PROPOSITION 8. The proof is essentially the same as for Proposition 7. M_n can be chosen so that $k_n \ln n = o(M_n)$, $M_n = o(\frac{n\tau_n^2}{k_n})$, and $M_n = o(\frac{n\tau_n^4}{k_n^2})$. These last two conditions are the ones needed to obtain the same upper bounds as in (20). □

6. Proofs.

6.1. *Proof of Theorem 1.* In the present setting all distributions are explicit and admit known densities with respect to the corresponding Lebesgue measure. We decompose any $y \in \mathbb{R}^n$ in two orthogonal components $y = \Phi\theta_y + y'$, with $\Phi^T y' = 0$. Then

$$dP_\theta(y) = c_1 \exp\left\{-\frac{1}{2\sigma_n^2}(\|\Phi\theta\|^2 + \|\Phi\theta_y\|^2 + \|y'\|^2 - 2\theta^T\Phi^T\Phi\theta_y)\right\},$$

$$d\widetilde{W}(\theta) = c_2 \exp\left\{-\frac{1}{2\tau_n^2}\|\Phi\theta\|^2\right\},$$

$$dP_\theta(y)\, d\widetilde{W}(\theta) = c_1 c_2 \exp\left\{-\frac{\sigma_n^2+\tau_n^2}{2\sigma_n^2\tau_n^2}\left\|\Phi\left(\theta - \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_y\right)\right\|^2 - \frac{1}{2(\sigma_n^2+\tau_n^2)}\|\Phi\theta_y\|^2 - \frac{1}{2\sigma_n^2}\|y'\|^2\right\},$$

where $c_1 = (2\pi)^{-n/2}\sigma_n^{-n}$ and $c_2 = (2\pi)^{-k_n/2}\tau_n^{-k_n}\det(\Phi^T\Phi)^{-1}$.

Using the Bayes rule, we get the density of $\widetilde{W}(d\theta|Y)$, in which we recognize the normal distribution

$$\widetilde{W}(d\theta|Y) = \mathcal{N}\left(\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_Y, \frac{\sigma_n^2\tau_n^2}{\sigma_n^2+\tau_n^2}(\Phi^T\Phi)^{-1}\right). \tag{21}$$

So we have an exact expression for $\widetilde{W}(d\theta|Y)$, but the centering and the variance do not correspond to the limit distribution given in Theorem 1. Therefore, we make use of the triangle inequality, with intermediate distribution $Q = \mathcal{N}(\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})$:

$$\begin{aligned}\|\widetilde{W}(d\theta|Y) - \mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \\ \le \|\widetilde{W}(d\theta|Y) - Q\|_{\mathrm{TV}} + \|Q - \mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}}.\end{aligned} \tag{22}$$

We first deal with the change in the variance, that is, the first term on the right in (22).

Let $\alpha_n = \frac{\tau_n}{\sigma_n}\sqrt{\ln(1+\frac{\sigma_n^2}{\tau_n^2})}$, and f and g be, respectively, the density functions of $\mathcal{N}(0, I_{k_n})$ and $\mathcal{N}(0, \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}I_{k_n})$. Let U be a random variable following

the chi-square distribution with k_n degrees of freedom $\chi^2(k_n)$. Let $\sqrt{\Phi^T\Phi}$ be a square root of the matrix $\Phi^T\Phi$. The total variation norm is invariant under the bijective affine map $\theta \mapsto \frac{1}{\sigma_n}\sqrt{\Phi^T\Phi}(\theta - \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_Y)$, so

$$\begin{aligned}\|\widetilde{W}(d\theta|Y) - Q\|_{\mathrm{TV}} &= \left\|\mathcal{N}(0, I_{k_n}) - \mathcal{N}\left(0, \frac{\tau_n^2}{\sigma_n^2+\tau_n^2} I_{k_n}\right)\right\|_{\mathrm{TV}} \\ &= \int_{\mathbb{R}^{k_n}} (g-f)_+ = \int_{\|x\| \le \sqrt{k_n}\alpha_n} (g(x) - f(x))\, d^n x \\ &= P\left(U \le \frac{\sigma_n^2+\tau_n^2}{\tau_n^2} k_n \alpha_n^2\right) - P(U \le k_n \alpha_n^2) \\ &= P\left(\sqrt{k_n}(\alpha_n^2 - 1) \le \frac{U-k_n}{\sqrt{k_n}} \le \sqrt{k_n}\left(\frac{\sigma_n^2+\tau_n^2}{\tau_n^2}\alpha_n^2 - 1\right)\right).\end{aligned}$$

As n goes to infinity, $\frac{U-k_n}{\sqrt{k_n}}$ converges toward $\mathcal{N}(0,1)$ in distribution. Using the Taylor expansion of ln, we find

$$\alpha_n^2 = 1 - \frac{\sigma_n^2}{2\tau_n^2} + o\left(\frac{\sigma_n^2}{\tau_n^2}\right)$$

and, therefore,

$$\begin{aligned}\sqrt{k_n}(\alpha_n^2 - 1) &\sim -\sqrt{k_n}\frac{\sigma_n^2}{2\tau_n^2}, \\ \sqrt{k_n}\left(\frac{\sigma_n^2+\tau_n^2}{\tau_n^2}\alpha_n^2 - 1\right) &\sim \sqrt{k_n}\frac{\sigma_n^2}{2\tau_n^2}.\end{aligned}$$

Since $k_n = o(\tau_n^4/\sigma_n^4)$, both these quantities go to 0. As a consequence, $\|\widetilde{W}(d\theta|Y) - Q\|_{\mathrm{TV}}$ goes to zero as n goes to infinity.

Let us now deal with the centering term, that is, the second term on the right in (22).

LEMMA 4. *Let U be a standard normal random variable, let $k \ge 1$ and let $Z \in \mathbb{R}^k$. Then*

$$\|\mathcal{N}(0, I_k) - \mathcal{N}(Z, I_k)\|_{\mathrm{TV}} = P(|U| \le \|Z\|/2) \le \|Z\|/\sqrt{2\pi}.$$

PROOF. Let g be the density of $\mathcal{N}(0, I_k)$. Then

$$\begin{aligned}\|\mathcal{N}(0, I_k) - \mathcal{N}(Z, I_k)\|_{\mathrm{TV}} &= \int_{\mathbb{R}^k} (g(x) - g(x-Z))_+\, d^k x \\ &= \int_{\{2x^T Z \le \|Z\|^2\}} (g(x) - g(x-Z))\, d^k x \\ &= P(U \le \|Z\|/2) - P(U + \|Z\| \le \|Z\|/2) \\ &\le \|Z\|/\sqrt{2\pi}.\end{aligned}$$

The last line comes from the density of $\mathcal{N}(0,1)$ being bounded by $1/\sqrt{2\pi}$. □

Using again the invariance of the total variation norm under the bijective affine map $\theta \mapsto \frac{1}{\sigma_n}\sqrt{\Phi^T\Phi}(\theta - \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_Y)$,

$$\begin{aligned}\|\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}) - Q\|_{\mathrm{TV}} &= \left\|\mathcal{N}(0, I_{k_n}) - \mathcal{N}\left(\frac{\sigma_n\sqrt{\Phi^T\Phi}\theta_Y}{\tau_n^2+\sigma_n^2}, I_{k_n}\right)\right\|_{\mathrm{TV}} \\ &\le \frac{1}{\sqrt{2\pi}}\frac{\sigma_n}{(\tau_n^2+\sigma_n^2)}\|\Phi\theta_Y\| \\ &\le \frac{1}{\sqrt{2\pi}}\frac{\sigma_n}{(\tau_n^2+\sigma_n^2)}(\|F_0\| + \sqrt{\varepsilon^T\Sigma\varepsilon}).\end{aligned}$$

$\varepsilon^T\Sigma\varepsilon$ is a random variable following $\sigma_n^2\chi^2(k_n)$ distribution. By Jensen's inequality, $E[\sqrt{\varepsilon^T\Sigma\varepsilon}] \le \sqrt{E[\varepsilon^T\Sigma\varepsilon]} = \sigma_n\sqrt{k_n}$. Therefore,

$$E\|\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}) - Q\|_{\mathrm{TV}} \le \frac{1}{\sqrt{2\pi}}\frac{\sigma_n}{\tau_n^2+\sigma_n^2}(\|F_0\| + \sigma_n\sqrt{k_n}),$$

which goes to zero under the assumptions of Theorem 1.

To conclude the proof, note that we deduce the results on $W(dF|Y)$ from the ones on $\widetilde{W}(d\theta|Y)$, by the linear relation $F = \Phi\theta$.

6.2. *Proof of Theorem 2.* We make the proof for $\widetilde{W}(d\theta|Y)$. Then the result for $W(dF|Y)$ is immediate. Our method is adapted from Boucheron and Gassiat (2009).

To any probability measure P on $\mathbb{R}^{k_n}$, we associate the probability

$$P^M = \frac{P(\cdot \cap \mathcal{E}_{\theta_0,\Phi}(M))}{P(\mathcal{E}_{\theta_0,\Phi}(M))} \tag{23}$$

with support in $\mathcal{E}_{\theta_0,\Phi}(M)$. It can be easily checked that

$$\|P - P^M\|_{\mathrm{TV}} = P(\mathcal{E}^c_{\theta_0,\Phi}(M)). \tag{24}$$

The proof is divided into three steps based on the use of M_n as a threshold to truncate the probability distributions. Lemma 5 below controls $E\|\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}) - \mathcal{N}^{M_n}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}}$, Lemma 6 controls $E\|\widetilde{W}^{M_n}(d\theta|Y) - \mathcal{N}^{M_n}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}}$ and Proposition 9 controls $E\|\widetilde{W}(d\theta|Y) - \widetilde{W}^{M_n}(d\theta|Y)\|_{\mathrm{TV}}$. Taken together, these results give Theorem 2.

LEMMA 5. *If $k_n < 4M_n$, then*

$$E\|\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}) - \mathcal{N}^{M_n}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \le 2e^{-(\sqrt{M_n}-2\sqrt{k_n})^2/8}.$$

If $k_n = o(M_n)$, for n large enough, this bound can be replaced by $e^{-M_n/9}$.

PROOF OF LEMMA 5. To control this quantity, we consider two cases, depending on whether θ_Y is near or far from θ_0:

$$\begin{aligned}
&\|\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}) - \mathcal{N}^{M_n}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \\
&\quad = \mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})(\mathcal{E}^c_{\theta_0,\Phi}(M_n)) \\
&\quad \le \mathbb{1}_{(\theta_Y-\theta_0)^T\Phi^T\Phi(\theta_Y-\theta_0)>\sigma_n^2 M_n/4} \\
&\qquad + \mathcal{N}(\theta_0, \sigma_n^2(\Phi^T\Phi)^{-1})(\mathcal{E}^c_{\theta_0,\Phi}(M_n/4)).
\end{aligned} \tag{25}$$

Let U be a random variable following a $\chi^2(k_n)$ distribution. Taking the expectation on both sides of (25) gives

$$E\|\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}) - \mathcal{N}^{M_n}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \le 2P(U > M_n/4).$$

Now, Cirelson's inequality [see, e.g., Massart (2007)]

$$P(\sqrt{U} > \sqrt{k_n} + \sqrt{2x}) \le \exp(-x) \tag{26}$$

used with $x = \frac{(\sqrt{M_n}-2\sqrt{k_n})^2}{8}$ implies Lemma 5. $\square$

LEMMA 6. *If* $\sup_{\|\Phi h\|^2 \le \sigma_n^2 M_n, \|\Phi g\|^2 \le \sigma_n^2 M_n} \frac{w(\theta_0+h)}{w(\theta_0+g)} \to 1$ *as* $n \to \infty$, *then*

$$E\|\widetilde{W}^{M_n}(d\theta|Y) - \mathcal{N}^{M_n}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})\|_{\mathrm{TV}} \to 0 \qquad \text{as } n \to \infty.$$

PROOF. Let us first note that, for every θ and τ in $\mathbb{R}^{k_n}$, for every $Y \in \mathbb{R}^n$,

$$\begin{aligned}
\frac{dP_\theta(Y)}{dP_\tau(Y)} &= \exp\left\{\frac{-\|\Phi\theta\|^2 + \|\Phi\tau\|^2 - 2Y^T\Phi(\tau-\theta)}{2\sigma_n^2}\right\} \\
&= \frac{d\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})(\theta)}{d\mathcal{N}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})(\tau)}.
\end{aligned} \tag{27}$$

This directly comes from the expressions for the Gaussian densities.

In the following the first lines are just rewriting. Then we use Jensen's inequality with the convex function $x \mapsto (1-x)_+$, and make use of (27). We abbreviate $\mathcal{N}^{M_n}(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1})$ into $\mathcal{N}^{M_n}$:

$$\begin{aligned}
&\|\widetilde{W}^{M_n}(d\theta|Y) - \mathcal{N}^{M_n}\|_{\mathrm{TV}} \\
&\quad = \int \left(1 - \frac{d\mathcal{N}^{M_n}(\theta)}{d\widetilde{W}^{M_n}(\theta|Y)}\right)_+ d\widetilde{W}^{M_n}(\theta|Y) \\
&\quad = \int \left(1 - \frac{d\mathcal{N}^{M_n}(\theta)\int (w(\tau)/d\mathcal{N}^{M_n}(\tau))\, dP_\tau(Y)\, d\mathcal{N}^{M_n}(\tau)}{w(\theta)\, dP_\theta(Y)}\right)_+ d\widetilde{W}^{M_n}(\theta|Y)
\end{aligned}$$

$$\leq \int\int \left(1 - \frac{w(\tau)\,d\mathcal{N}^{M_n}(\theta)\,dP_\tau(Y)}{w(\theta)\,d\mathcal{N}^{M_n}(\tau)\,dP_\theta(Y)}\right)_+ d\mathcal{N}^{M_n}(\tau)\,d\widetilde{W}^{M_n}(\theta|Y)$$

$$= \int\int \left(1 - \frac{w(\tau)}{w(\theta)}\right)_+ d\mathcal{N}^{M_n}(\tau)\,d\widetilde{W}^{M_n}(\theta|Y)$$

$$\leq 1 - \inf_{\|\Phi h\|^2 \leq \sigma_n^2 M_n, \|\Phi g\|^2 \leq \sigma_n^2 M_n} \frac{w(\theta_0 + h)}{w(\theta_0 + g)}. \qquad \square$$

PROPOSITION 9 (Posterior concentration). *Suppose that conditions (1), (2) and (3) of Theorem 2 hold. Then*

$$E\|\widetilde{W}(d\theta|Y) - \widetilde{W}^{M_n}(d\theta|Y)\|_{\mathrm{TV}} = E[\widetilde{W}(\mathcal{E}^C_{\theta_0,\Phi}(M_n)|Y)] \to 0 \qquad \text{as } n \to \infty.$$

Proposition 9 is proved in Appendix A in the supplemental article [Bontemps (2011)]. However, we state here the following important lemma, because of its significance.

LEMMA 7. *Let $a \in \mathbb{R}^n$ such that $\Phi^T a = 0$. Then, for any $y \in \mathbb{R}^n$, $W(dF|Y = y) = W(dF|Y = y + a)$.*

Lemma 7 states that the distribution $W(dF|Y)$ is invariant under any translation of Y orthogonal to $\langle\phi\rangle$. Now, regard $W(dF|Y)$ as a random variable. Then any statement on $W(dF|Y)$ or $\widetilde{W}(d\theta|Y)$ valid when $Y \sim \mathcal{N}(F_0, \sigma_n^2 I_n)$ with $F_0 \in \langle\phi\rangle$ can be extended at zero cost by Lemma 7 to the case $F_0 \in \mathbb{R}^n$. For instance, proving Proposition 9 in the case $F_0 = \Phi\theta_0$ is enough.

6.3. *Proof of Theorem 3.* We begin with (13). Consider the following Taylor expansion:

$$\begin{aligned} G(F) - G(Y_{\langle\phi\rangle}) &= \dot{G}_{F_{\langle\phi\rangle}}(F - Y_{\langle\phi\rangle}) \\ &\quad + \frac{1}{2}\int_0^1 (1-t) D^2_{F_{\langle\phi\rangle} + t(F - F_{\langle\phi\rangle})} G(F - F_{\langle\phi\rangle}, F - F_{\langle\phi\rangle})\,dt \\ &\quad - \frac{1}{2}\int_0^1 (1-t) D^2_{F_{\langle\phi\rangle} + t(Y_{\langle\phi\rangle} - F_{\langle\phi\rangle})} G(Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}, Y_{\langle\phi\rangle} - F_{\langle\phi\rangle})\,dt \end{aligned}$$

using the Lagrange form of the error term. Suppose that $F \in \langle\phi\rangle$, $\|F - F_{\langle\phi\rangle}\|^2 \leq \sigma_n^2 M_n$ and $\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 \leq \sigma_n^2 M_n$. Then, for any $b \in \mathbb{R}^p$,

$$|b^T(G(F) - G(Y_{\langle\phi\rangle}) - \dot{G}_{F_{\langle\phi\rangle}}(F - Y_{\langle\phi\rangle}))| \leq \|b\| B_{F_{\langle\phi\rangle}}(M_n).$$

On the other hand, $\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b} \geq \sqrt{\|\Gamma^{-1}_{F_{\langle\phi\rangle}}\|^{-1}}\|b\|$. Moreover,

$$\left\| W\left(d\frac{b^T\dot{G}_{F_{\langle\phi\rangle}}(F - Y_{\langle\phi\rangle})}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}}\Big| Y\right) - \mathcal{N}(0,1)\right\|_{\mathrm{TV}}$$
$$\leq \|W(dF|Y) - \mathcal{N}(Y_{\langle\phi\rangle}, \sigma_n^2\Sigma)\|_{\mathrm{TV}}.$$

Let $\eta_n = \sqrt{\|\Gamma^{-1}_{F_{\langle\phi\rangle}}\|}B_{F_{\langle\phi\rangle}}(M_n)$, which tends to 0 by hypothesis. Let also

$$I_{\eta_n} = \{x \in \mathbb{R} : \exists x' \in I, |x - x'| \leq \eta_n\}.$$

Note that $\psi(I_{\eta_n}) \leq \psi(I) + \sqrt{\frac{2}{\pi}}\eta_n$.

Gathering all this information, we can get the upper bound

$$W\left(\frac{b^T(G(F) - G(Y_{\langle\phi\rangle}))}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \in I\Big| Y\right)$$
$$\leq W\left(\frac{b^T\dot{G}_{F_{\langle\phi\rangle}}(F - Y_{\langle\phi\rangle})}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \in I_{\eta_n}\Big| Y\right)$$
$$+ \mathbb{1}_{\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n} + W(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n|Y)$$
$$\leq \psi(I) + \sqrt{\frac{2}{\pi}}\eta_n + \|W(dF|Y) - \mathcal{N}(Y_{\langle\phi\rangle}, \sigma_n^2\Sigma)\|_{\mathrm{TV}}$$
$$+ \mathbb{1}_{\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n} + W(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n|Y).$$

A lower bound is obtained in the same way. Taking the expectation,

$$E\left|W\left(\frac{b^T(G(F) - G(Y_{\langle\phi\rangle}))}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \in I\Big| Y\right) - \psi(I)\right|$$
$$\leq o(1) + P(\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n) \tag{28}$$
$$+ E[W(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n|Y)].$$

But $\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2$ follows the $\sigma_n^2\chi^2(k_n)$ distribution, and since $k_n = o(M_n)$,

$$P(\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n) = o(1).$$

To bound (28), we use the following:

LEMMA 8. *Suppose that the conditions of either Theorems 1 or 2 are satisfied. Then*

$$E[W(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n|Y)] \to 0 \qquad \text{as } n \to \infty.$$

PROOF. For smooth priors, this is an immediate corollary of Proposition 9. Let us suppose we are under the conditions of Theorem 1.

Let Z be a $\mathcal{N}(0, \frac{\sigma_n^2\tau_n^2}{\sigma_n^2+\tau_n^2}\Sigma)$ random vector in $\mathbb{R}^n$ independent on Y, and U a random variable following $\chi^2(k_n)$. From (21) we get

$$W(dF|Y) = \mathcal{N}\left(\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}Y_{\langle\phi\rangle}, \frac{\sigma_n^2\tau_n^2}{\sigma_n^2+\tau_n^2}\Sigma\right).$$

Therefore,

$$\begin{aligned}
&W(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n|Y)\\
&\quad= P\left(\left\|Z + \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\|^2 > \sigma_n^2 M_n\right)\\
&\quad\le P\left(\|Z\| > \sigma_n\sqrt{M_n} - \left\|\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\|\right)\\
&\quad\le \begin{cases} 1, & \text{if } \left\|\dfrac{\tau_n^2}{\sigma_n^2+\tau_n^2}Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\| > \dfrac{2\sigma_n\sqrt{M_n}}{3},\\ P\left(\|Z\|^2 > \sigma_n^2\dfrac{M_n}{9}\right) = P\left(U > \dfrac{\sigma_n^2+\tau_n^2}{\tau_n^2}\dfrac{M_n}{9}\right), & \text{otherwise.}\end{cases}
\end{aligned}$$

Since $k_n = o(M_n)$, $P(U > M_n/9) = o(1)$. On the other hand,

$$\begin{aligned}
\left\|\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\| &= \left\|\Sigma\left(\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\varepsilon + \frac{\sigma_n^2}{\sigma_n^2+\tau_n^2}F_0\right)\right\|\\
&\le \|\Sigma\varepsilon\| + \frac{\sigma_n}{\sqrt{\sigma_n^2+\tau_n^2}}\|F_0\|.
\end{aligned}$$

Since $\|F_0\| = o(\tau_n^2/\sigma_n)$, $\frac{\sigma_n^2\|F_0\|^2}{\sigma_n^2+\tau_n^2} = o(1) < \frac{M_n}{9}$ for n large enough. $\|\Sigma\varepsilon\|^2$ is a $\sigma_n^2\chi^2(k_n)$ variable. Therefore, for n large enough,

$$E[W(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n|Y)] \le 2P(U > M_n/9) = o(1). \qquad \square$$

Now, (28) gives (13).

The proof of the frequentist assertion (14) is similar and delayed to Appendix C in the supplemental article [Bontemps (2011)].

Acknowledgments. The author would like to thank E. Gassiat and I. Castillo for valuable discussions and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors" (DOI: 10.1214/11-AOS912SUPP;

.pdf). This contains the proofs of various technical results stated in the main article "Bernstein–von Mises Theorems for Gaussian regression with increasing number of regressors."

REFERENCES

Bontemps, D. (2011). Supplement to "Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors." DOI:10.1214/11-AOS912SUPP.

Boucheron, S. and Gassiat, E. (2009). A Bernstein–von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148. MR2471588

Castillo, I. (2010). A semi-parametric Bernstein–von Mises theorem. *Probab. Theory Related Fields*. DOI:10.1007/s00440-010-0316-5.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **36** 453–471. MR1053841

Clarke, B. and Ghosal, S. (2010). Reference priors for exponential families with increasing dimension. *Electron. J. Stat.* **4** 737–780. MR2678969

de Boor, C. (1978). *A Practical Guide to Splines*. *Applied Mathematical Sciences* **27**. Springer, New York. MR0507062

Freedman, D. (1999). Wald lecture: On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140.

Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* **5** 315–331. MR1681701

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** 49–68. MR1790613

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

Kim, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700. MR2283713

Kim, Y. and Lee, J. (2004). A Bernstein–von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32** 1492–1512. MR2089131

Massart, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

Rivoirard, V. and Rousseau, J. (2009). Bernstein von Mises theorem for linear functionals of the density. Available at `http://arxiv.org/abs/0908.4167`.

Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.* **97** 222–235. MR1947282

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

Tsybakov, A. B. (2004). *Introduction à L'estimation Non-Paramétrique*. Springer, Berlin.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge.

Laboratoire de Mathématiques d'Orsay
Université Paris-Sud
UMR8628, Bât. 425
F-91405, Orsay
France
E-mail: dominique.bontemps@gmail.com