Electronic Journal of Statistics
Vol. 7 (2013) 217–243
ISSN: 1935-7524
DOI: 10.1214/13-EJS768

Semiparametric Bernstein–von Mises for the error standard deviation

René de Jonge

Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands
e-mail: r.d.jonge@tue.nl

and

Harry van Zanten

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
P.O. Box 94248
1098 GE Amsterdam
The Netherlands
e-mail: hvzanten@uva.nl

Abstract: We study Bayes procedures for nonparametric regression problems with Gaussian errors, giving conditions under which a Bernstein–von Mises result holds for the marginal posterior distribution of the error standard deviation. We apply our general results to show that a single Bayes procedure using a hierarchical spline-based prior on the regression function and an independent prior on the error variance, can simultaneously achieve adaptive, rate-optimal estimation of a smooth, multivariate regression function and efficient, $\sqrt{n}$-consistent estimation of the error standard deviation.

AMS 2000 subject classifications: Primary 62G09; secondary 62C10, 62G20.
Keywords and phrases: Nonparametric regression, Bayesian inference, estimation of error variance, semiparametric Bernstein-von Mises.

Received May 2012.

1. Introduction

In this paper we study the asymptotic behavior of the marginal posterior for the error standard deviation in a nonparametric, fixed design regression model with Gaussian errors. We suppose we have observations $Y_1, \dots, Y_n$ satisfying

$$Y_i = f_0(x_i) + \sigma_0 Z_i, \quad i = 1, \dots, n,$$

[*]Research supported by the Netherlands Organization for Scientific Research (NWO).

where $x_1, \ldots, x_n$ are known elements of a general design space $\mathcal{X}$, the variables $Z_1, \ldots, Z_n$ are independent, standard normal and both the regression function $f_0 : \mathcal{X} \to \mathbb{R}$ and the error standard deviation $\sigma_0 > 0$ are unknown. We can then make Bayesian inference about the parameters f and σ by endowing them with independent priors Π_f and Π_σ, respectively, and considering the resulting posterior distribution $\Pi(\cdot \,|\, Y_1, \ldots, Y_n)$. We study the asymptotic behavior of the marginal posterior distribution $B \mapsto \Pi(\sigma \in B \,|\, Y_1, \ldots, Y_n)$ of the parameter σ for $n \to \infty$.

Although in most cases the main interest is in estimating the regression function f, making accurate inference about the error variance σ can also be important. In regression analysis it is common to report an estimate of σ to quantify the magnitude of the measurement errors in the data or to assess model fit. It is quite natural to attempt to estimate σ in an efficient way. In the frequentist literature this problem has been studied for a long time and in increasing generality. See for instance the recent paper Brown and Levine (2007) for historical comments and rather extensive references. The efficient estimation of the error standard deviation or variance in nonparametric regression has so far received little attention in the Bayesian literature however. The existing theorems focus on contraction rates for the posterior distribution of the regression function f and at best give only crude rates for the posterior distribution of σ. Theorems about the asymptotic shape of the posterior of σ have not been obtained so far.

The general rate of contraction result for fixed design regression obtained in Ghosal and Van der Vaart (2007) gives conditions under which the posterior for the regression function f contracts around the true f_0 at a certain rate ε_n as $n \to \infty$, under the assumption that σ_0 is known. As has been observed several times in the literature (see e.g. Van der Vaart and Van Zanten (2008a), Van der Vaart and Van Zanten (2009), De Jonge and Van Zanten (2010)) this result can be extended to the case that σ_0 is unknown, see also the appendix to this paper. In that case one also obtains a rate for the marginal posterior of σ. Specifically, the (extended version of) existing general results give conditions under which, for a given sequence $\varepsilon_n \to 0$, it holds that

$$\Pi\Big((f,\sigma) : \frac{1}{n}\sum_{i=1}^{n}(f-f_0)^2(x_i) + |\sigma-\sigma_0|^2 \geq M^2\varepsilon_n^2 \,|\, Y_1, \ldots, Y_n\Big) \overset{P_0}{\to} 0 \qquad (1.1)$$

as $n \to \infty$, for every sufficiently large $M > 0$. Here the convergence is in probability under the true model.

A result like (1.1) implies in particular that the marginal posterior for σ is asymptotically concentrated on an interval with length of the order ε_n around the true value σ_0. Since ε_n is also a bound for the rate of contraction of the marginal posterior for f however, it is a "nonparametric rate" that will be slower than the parametric rate $n^{-1/2}$ if the space of regression functions that are considered is infinite-dimensional. The rate bound ε_n for the one-dimensional parameter σ is therefore typically very crude and it is natural to ask whether in fact the actual rate of contraction for the marginal posterior for σ can be faster than the rate for the regression function f.

In the extreme case that the regression function f is completely known and σ is the only unknown parameter in the problem, the classical Bernstein von–Mises (BvM) theorem asserts that under minimal regularity conditions, the posterior distribution of σ contracts around the true value σ_0 at the rate $n^{-1/2}$. Moreover, it says that the posterior law of $\sqrt{n}(\sigma - \sigma_0)$ behaves asymptotically like a normal distribution $N(\Delta_n, I_{\sigma_0}^{-1})$, with Δ_n a sequence of random variables with an asymptotic $N(0, I_{\sigma_0}^{-1})$-distribution under $\mathbb{P}_0$ and I_{σ_0} the Fisher information for σ_0. The precise statement is recalled in the next section. The BvM result implies in particular that the posterior for σ correctly quantifies the uncertainty about the parameter. Specifically, if credible bounds $l_n < u_n$ are determined such that for a fixed level $\alpha \in (0,1)$ it holds that $\Pi(\sigma \in (l_n, u_n) \,|\, Y_1, \ldots, Y_n) \geq 1 - \alpha$, then the BvM theorem implies that the credible interval (l_n, u_n) also has frequentist coverage probability $1 - \alpha$ asymptotically, i.e. $\liminf_{n\to\infty} \mathbb{P}_0(\sigma_0 \in (l_n, u_n)) \geq 1 - \alpha$. Moreover, the length $u_n - l_n$ of the credible interval asymptotically coincides with the length of an optimal confidence interval. We refer to the discussion in Section 1.5 of Castillo (2012a) for more details.

In this paper we investigate if and how this changes if the regression function f is unknown. In this case we know that (1.1) holds for instance if $f_0 \in \mathcal{F}$ and $\sigma_0 \in [a, b]$, say, and we place independent priors Π_f and Π_σ on $\mathcal{F}$ and $[a, b]$, respectively, Π_σ having a positive, continuous Lebesgue density and Π_f such that for positive numbers $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \leq \varepsilon_n$ and constants $c_1, c_2 > 0$ it holds that for every $c_3 > 1$, there exist measurable subsets $\mathcal{F}_n \subset \mathcal{F}$ and a constant $c_4 > 0$ such that

$$\Pi_f(f : \|f - f_0\|_n \leq \tilde{\varepsilon}_n) \geq c_1 e^{-c_2 n \tilde{\varepsilon}_n^2}, \tag{1.2}$$

$$\Pi_f(\mathcal{F} \backslash \mathcal{F}_n) \leq e^{-c_3 n \tilde{\varepsilon}_n^2}, \tag{1.3}$$

$$\log N(\bar{\varepsilon}_n, \mathcal{F}_n, \|\cdot\|_n) \leq c_4 n \bar{\varepsilon}_n^2. \tag{1.4}$$

Here $\|g\|_n^2 = n^{-1} \sum g^2(x_i)$ and for a metric space (A, d) and $\varepsilon > 0$, $N(\varepsilon, A, d)$ is the minimal number of balls of d-radius ε needed to cover A. (See Theorem A.1 in the appendix for this result.) We prove below (see Theorem 2.2) that if in addition $n\varepsilon_n^4 \to 0$ and

$$\int_0^{a\varepsilon_n} \sqrt{\log N(\delta, \mathcal{F}_n, \|\cdot\|_n)}\, d\delta \to 0 \quad \text{for all } a > 0, \tag{1.5}$$

then the BvM assertion holds for the marginal posterior distribution of σ. In particular, the marginal posterior distribution of σ then has the same, optimal asymptotic behavior as in the case that f is known.

In the literature various papers can be found that deal with the verification of conditions (1.2)–(1.4) for specific families of priors on f. See for instance Ghosal and Van der Vaart (2007), Van der Vaart and Van Zanten (2008a), Van der Vaart and Van Zanten (2009), De Jonge and Van Zanten (2010), De Jonge and Van Zanten (2012), Tokdar (2011), Bhattacharya, Pati and Dunson (2012). These results can however not be applied directly to verify also the additional condition (1.5). The reason is that in the cited papers, the constructed sieves $\mathcal{F}_n$ that verify (1.3) and (1.4), are typically too large for condition (1.5) to hold. Therefore, verifying the conditions of our general BvM theorem for a specific prior usually involves the careful construction of alternative sieves. The new, smaller sieves should be such that the remaining mass condition (1.3) is still fulfilled and in addition the entropy $\log N(\delta, \mathcal{F}_n, \|\cdot\|_n)$ can be controlled for arbitrarily small δ, so that (1.5) can be verified.

In this paper we carry out this task for Gaussian process priors and for a spline-based prior on the regression function. In the case of Gaussian process priors, it is known that conditions (1.2)–(1.4) can be replaced by single condition on the so-called concentration function of the prior, cf. Van der Vaart and Van Zanten (2008a). Roughly speaking we prove in Theorem 3.1 below that if in addition to this condition the rate ε_n is fast enough and the sample paths of the Gaussian prior have regularity larger than $d/2$, for d the dimension of the covariate space, then the BvM statement holds. We give details for two specific popular families of Gaussian priors: multiply integrated Brownian motions and the class of Matérn processes. In both cases we find that BvM holds if the prior is rough enough relative to the degree of smoothness of the true regression function f_0. In some generality it is known that if we want optimal contraction rates for f using a Gaussian prior, then the regularities of the truth and the prior should be equal (see Van der Vaart and Van Zanten (2008a), Castillo (2008)). In the examples we work out we find that for BvM for σ to hold it is not necessary that the smoothnesses are matched exactly however. Some degree of oversmoothing is allowed and an arbitrary degree of undersmoothing, cf. Section 3.2. In particular, the rate of contraction of the marginal posterior for f may be sub-optimal, while still having an optimal asymptotic behavior of the posterior for σ. This is in line with the findings of Castillo (2012a) in the context of the white noise model.

The second type of concrete priors we study are spline-based priors studied before in De Jonge and Van Zanten (2012). More precisely, we consider a hierarchical prior on functions on $[0,1]^d$, defined structurally as a spline of fixed order, with randomly placed, regularly spaced knots and random B-spline coefficients (details in Section 4). In De Jonge and Van Zanten (2012) it was shown that when properly constructed, such a prior yields adaptive, nearly rate-optimal estimation of a smooth regression function f. We investigate this prior in this paper because we are interested in the question whether or not we can have adaptive estimation of f and BvM for σ at the same time. In Theorem 4.1 we show, by constructing appropriate sieves, that this is indeed possible. For the spline prior we prove that if the true f_0 is a d-variate function with (Hölder-) regularity β, then BvM for σ holds if $\beta > d$. So in that case we have a single procedure that yields both efficient estimation of the error standard deviation and adaptive, nearly rate-optimal estimation of f across a range of regularities. The specific priors that we analyze are Gaussian or conditionally Gaussian. This is technically convenient, since it allows us to use tools from Gaussian process theory. However we stress that our general BvM theorems are valid outside the Gaussian realm as well.

Our general result can be viewed as a semiparametric Bernstein-von Mises theorem. In general, semiparametric BvM theorems deal with the asymptotic behavior of posterior distributions of finite-dimensional parameters in the presence of an infinite-dimensional "nuisance" parameter. Theorems of this type have recently been established by several authors, see for instance Shen (2002), De Blasi and Hjort (2009), Castillo (2012a), Bickel and Kleijn (2012), Rivoirard and Rousseau (2012). Our problem in fact fits into the general framework of Castillo (2012a) (up to minor adaptations) and we will use his results to derive our BvM theorem for the error standard deviation.

The remainder of the paper is organized as follows. After recalling the parametric BvM theorem in Section 2.1 we present our general semiparametric results for the error standard deviation in Section 2.2. In Section 3 we consider the special case that the prior on f is Gaussian. We formulate a general theorem and verify the conditions for the two particular examples mentioned above. Section 4 treats the hierarchical spline-based priors. We prove that they yield simultaneous adaptation for f and BvM for σ. The proof of our general theorem is given in Section 5. In the appendix, which we added for the sake of completeness, we state and prove a theorem giving sufficient conditions for the contraction rate result (1.1). This result is essentially known, but a proof has never been published.

2. General result

2.1. Prelude: Parametric Bernstein–von Mises

The main result of this paper is a semiparametric Bernstein–von Mises (BvM) theorem for the error standard variance in a fixed design regression model. As a prelude we first consider the parametric case in which we observe variables $Y_1, \dots, Y_n$ satisfying

$$Y_i = f_0(x_i) + \sigma Z_i, \qquad i = 1, \dots, n,$$

for known covariates $x_i \in \mathcal{X}$ and standard normal random variables Z_i. We now assume that the regression function f_0 is *known*, so that the error standard deviation $\sigma > 0$ is the only unknown parameter. We denote its true value by σ_0. Observe that in this case we simply have a sample of size n from the $N(0, \sigma^2)$-distribution, given by $X_i = Y_i - f_0(x_i)$, $i = 1, \dots, n$.

The BvM theorem in a smooth, parametric i.i.d. model like this one is classical. As an illustration and to connect to the semiparametric case studied ahead we briefly explain it. Let p_σ be the marginal density of X_i, $\ell_\sigma(x) = \log p_\sigma(x)$, $\dot{\ell}_\sigma(x) = \partial \ell_\sigma(x)/\partial\sigma$ and $\ddot{\ell}_\sigma(x) = \partial \dot{\ell}_\sigma(x)/\partial\sigma$. Then a Taylor expansion gives

$$\ell_\sigma(x) - \ell_{\sigma_0}(x) \approx (\sigma - \sigma_0)\dot{\ell}_{\sigma_0}(x) + \frac{1}{2}(\sigma - \sigma_0)^2 \ddot{\ell}_{\sigma_0}(x).$$

By the law of large numbers the average $-n^{-1}\sum_{i=1}^n \ddot{\ell}_{\sigma_0}(X_i)$ converges almost surely to the Fisher information $I_{\sigma_0} = -\mathbb{E}_0 \ddot{\ell}_{\sigma_0}(X_1) = \mathbb{V}\mathrm{ar}_0 \dot{\ell}_{\sigma_0}(X_1)$. It follows that for the full log-likelihood we have the LAN approximation

$$\log \prod_{i=1}^{n} \frac{p_\sigma}{p_{\sigma_0}}(X_i) \approx -\frac{1}{2} I_{\sigma_0} \Big(n(\sigma - \sigma_0)^2 - 2\sqrt{n}(\sigma - \sigma_0)\Delta_n \Big),$$

where

$$\Delta_n = I_{\sigma_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\sigma_0}(X_i).$$

By the central limit theorem, we have the weak convergence $\Delta_n \Rightarrow N(0, I_{\sigma_0}^{-1})$ as $n \to \infty$.

If we now put a prior on $(0,\infty)$ with a Lebesgue density π which is positive and continuous at σ_0, then for the corresponding posterior we have, for a Borel subset $B \subset \mathbb{R}$,

$$\Pi(\sqrt{n}(\sigma - \sigma_0) \in B \,|\, Y_1, \dots, Y_n) = \frac{\int_{\sqrt{n}(\sigma-\sigma_0)\in B} \prod_{i=1}^{n} \frac{p_\sigma}{p_{\sigma_0}}(X_i)\pi(\sigma)\, d\sigma}{\int_{\mathbb{R}_+} \prod_{i=1}^{n} \frac{p_\sigma}{p_{\sigma_0}}(X_i)\pi(\sigma)\, d\sigma}.$$

By the LAN approximation, the integrands are approximately equal to a constant times

$$\pi(\sigma) \exp\Big(-\frac{1}{2} I_{\sigma_0} (\sqrt{n}(\sigma - \sigma_0) - \Delta_n)^2 \Big).$$

Making a change of variable $\sqrt{n}(\sigma - \sigma_0) = h$ we then see that the posterior probability that $\sqrt{n}(\sigma - \sigma_0)$ falls in the set B approximately equals $N(\Delta_n, I_{\sigma_0}^{-1})(B)$ for large n.

This somewhat loose argumentation can be made precise and it can be shown that in probability, the total variation distance between the posterior distribution of $\sqrt{n}(\sigma - \sigma_0)$ and the $N(\Delta_n, I_{\sigma_0}^{-1})$-distribution vanishes as $n \to \infty$, cf. e.g. Van der Vaart (1998). It is easily verified that in this case

$$\Delta_n = \frac{\sigma_0}{2\sqrt{n}} \sum_{i=1}^{n} (Z_i^2 - 1), \qquad I_{\sigma_0} = \frac{2}{\sigma_0^2}. \tag{2.1}$$

In the next section we state the semiparametric version of this result for the case that the regression function f is in fact unknown. It turns out that there is no loss of information (in the semiparametric sense) for the error standard deviation and that under relatively mild conditions on the prior for the nonparametric part f, the asymptotic behavior of the marginal posterior for $\sqrt{n}(\sigma - \sigma_0)$ is the same as if f were known.

2.2. Semiparametric Bernstein–von Mises

Now suppose that we have observations $Y_1, \dots, Y_n$ from the regression model

$$Y_i = f(x_i) + \sigma Z_i, \quad i = 1, \dots, n, \tag{2.2}$$

with fixed and known design points $x_1, \dots, x_n$ in the set $\mathcal{X}$, an *unknown* regression function $f : \mathcal{X} \to \mathbb{R}$, an unknown constant $\sigma > 0$, and with $Z_1, \dots, Z_n$ independent standard Gaussian random variables. We assume that the true parameter (σ_0, f_0) belongs to the set $(0, \infty) \times \mathcal{F}$, for $\mathcal{F}$ a measurable space of functions on $\mathcal{X}$. The corresponding true distribution of the data is denoted by $\mathbb{P}_0$.

The log-likelihood is given by

$$\ell_n(\sigma, f; Y_1, \dots, Y_n) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - f(x_i))^2.$$

We assume that for every n, the map $(\sigma, f, y) \mapsto \ell_n(\sigma, f; y_1, \dots, y_n)$ is a measurable map on $(0, \infty) \times \mathcal{F} \times \mathbb{R}^n$. Note that this is the case for instance if $\mathcal{X}$ is a topological space and $\mathcal{F}$ is a measurable subset of the space of $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$, endowed with its Borel sigma-field.

To make Bayesian inference about f and σ we endow the pair (σ, f) with a product prior distribution of the form $\Pi = \Pi_\sigma \times \Pi_f$. Here Π_σ is a distribution on $(0, \infty)$ with a positive and continuous Lebesgue density and Π_f is a distribution on $\mathcal{F}$. In view of the measurability assumptions the corresponding posterior distribution is well defined and given by Bayes' formula. For A and B measurable subsets of $(0, \infty)$ and $\mathcal{F}$, respectively, the posterior measure of the set $A \times B$ is denoted by $\Pi(A \times B \,|\, Y_1, \dots, Y_n)$ or $\Pi(\sigma \in A, f \in B \,|\, Y_1, \dots, Y_n)$.

The following theorem deals with the marginal posterior distribution of the parameter σ. It gives conditions under which we have, as in the case that f is known, that the posterior distribution of $\sqrt{n}(\sigma - \sigma_0)$ asymptotically behaves as an $N(\Delta_n, I_{\sigma_0}^{-1})$-distribution, where Δ_n and I_{σ_0} are as in (2.1). Note that we still have the weak convergence $\Delta_n \Rightarrow N(0, I_{\sigma_0}^{-1})$ under $\mathbb{P}_0$, by the central limit theorem.

The existing general contraction rate theorems for fixed design regression give conditions under which the posterior contracts around the true parameter (σ_0, f_0). More precisely, for a sequence of positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ they give conditions under which there exist measurable subsets $\mathcal{F}_n \subset \mathcal{F}$ such that

$$\Pi((\sigma, f) \in (0, \infty) \times \mathcal{F}_n : |\sigma - \sigma_0| + \|f - f_0\|_n \le \varepsilon_n \,|\, Y_1, \dots, Y_n) \overset{P_0}{\to} 1 \qquad (2.3)$$

as $n \to \infty$, where, as before, the norm $\|\cdot\|_n$ is the L^2-norm associated with the empirical measure on the design points, i.e. $\|g\|_n^2 = n^{-1} \sum g^2(x_i)$. (Since a full proof of this exact statement appears never to have been given in the literature, we provide it in the appendix of the paper for the sake of completeness. See Theorem A.1.) The case that σ_0 is known is covered by these general results as well. Following Castillo (2012a), we denote the posterior distribution for f in the model that σ_0 is known by $\Pi^{\sigma=\sigma_0}(\cdot \,|\, Y_1, \dots, Y_n)$. In this notation, the general theory gives conditions under which

$$\Pi^{\sigma=\sigma_0}(f \in \mathcal{F}_n : \|f - f_0\|_n \le \varepsilon_n \,|\, Y_1, \dots, Y_n) \overset{P_0}{\to} 1 \qquad (2.4)$$

as $n \to \infty$ (see e.g. Ghosal and Van der Vaart (2007)).

The rate ε_n should be viewed as the contraction rate that is achieved for the nonparametric part of the statistical problem. The following theorem states that if this rate is fast enough, namely $n\varepsilon_n^4 \to 0$, then under the additional entropy condition (1.5), we have the BvM result for the error standard deviation σ. The proof of the theorem is given in Section 5.

Theorem 2.1. *Consider positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ and $n\varepsilon_n^4 \to 0$. If there exist measurable subsets $\mathcal{F}_n \subset \mathcal{F}$ such that (2.3), (2.4) and (1.5) hold, then with Δ_n and I_{σ_0} given by (2.1) we have*

$$\sup_B \left| \Pi(\sqrt{n}(\sigma - \sigma_0) \in B, f \in \mathcal{F} | Y_1, \dots, Y_n) - N(\Delta_n, I_{\sigma_0}^{-1})(B) \right| \overset{P_0}{\to} 0$$

as $n \to \infty$, where the supremum is taken over all measurable subsets $B \subset \mathbb{R}$.

Existing general theorems give sufficient conditions on the prior Π_f for (2.3) and (2.4) to hold. Full proofs are only given in the literature for the case that σ is known (see Ghosal and Van der Vaart (2007)), which only takes care of (2.4). It has been noted however that these results can be adapted to deal with the case that σ_0 belongs to a known compact interval $[a, b]$ and Π_σ is a prior concentrated on $[a, b]$. For completeness, we give a precise result in Theorem A.1 in the appendix. Admittedly, the assumption that the standard deviation belongs to a compact interval is restrictive. Extending the general rate result given in the appendix to alleviate this restriction is therefore desirable, but is not completely straightforward. We note that our general theorem, Theorem 2.1, does not require σ to be in a compact set. Hence, a generalization of Theorem A.1 will immediately yield a generalization of the following theorem as well.

Theorem 2.2. *Suppose that $\sigma \in [a, b]$ and Π_σ is concentrated on $[a, b]$. Consider positive numbers $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \le \varepsilon_n$ such that $n(\tilde{\varepsilon}_n \wedge \bar{\varepsilon}_n)^2 \gtrsim \log n$ and $n\varepsilon_n^4 \to 0$. Suppose that for constants $c_1, c_2 > 0$ we have that for every $c_3 > 1$, there exist measurable subsets $\mathcal{F}_n \subset \mathcal{F}$ and a constant $c_4 > 0$ such that conditions (1.2)–(1.5) are fulfilled. Then with Δ_n and I_{σ_0} given by (2.1) we have*

$$\sup_B \left| \Pi(\sqrt{n}(\sigma - \sigma_0) \in B, f \in \mathcal{F} | Y_1, \dots, Y_n) - N(\Delta_n, I_{\sigma_0}^{-1})(B) \right| \overset{P_0}{\to} 0$$

as $n \to \infty$, where the supremum is taken over all measurable subsets $B \subset \mathbb{R}$.

Proof. Combining Theorems A.1 and 2.1 yields the result. □

In the next two sections we verify the conditions of Theorem 2.2 for two classes of priors Π_f: Gaussian process priors and hierarchical spline-based priors.

3. Gaussian process priors

3.1. General Gaussian priors

We now specialize to the case that $\mathcal{X} = [0, 1]^d$ for some $d \in \mathbb{N}$. As prior Π_f on the regression function f we employ the law of a Gaussian random element W in the space $C([0,1]^d)$ of continuous functions on $[0,1]^d$. We denote the reproducing kernel Hilbert space (RKHS) of W by $\mathbb{H}$. For $f_0 \in C([0,1]^d)$ the true regression function, the associated concentration function is denoted by φ_{f_0}, that is to say

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h-f_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon), \quad \varepsilon > 0. \tag{3.1}$$

(See the papers Van der Vaart and Van Zanten (2008a) and Van der Vaart and Van Zanten (2008b) and the references therein for these fundamental concepts.) As in Theorem 2.2, the error standard deviation is assumed to belong to $[a, b]$ and Π_σ is concentrated on that interval. The general theory for Gaussian process priors then says that if $\varepsilon_n \to 0$ is such that $n\varepsilon_n^2 \to \infty$ and

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2, \tag{3.2}$$

then the marginal posteriors for f and σ contract at the rate ε_n around their true values, cf. Theorem 3.3 of Van der Vaart and Van Zanten (2008a).

The theorem below essentially states that if in addition to (3.2) we have $n\varepsilon_n^4 \to 0$ and W has degree of regularity $\alpha > d/2$, then BvM holds true. Specifically, we shall assume that W takes values in the Hölder space $C^\gamma[0,1]^d$ for all $\gamma < \alpha$. (Recall that a function belongs to this space if for $\underline{\gamma}$ the largest integer strictly smaller than γ, it has continuous partial derivatives up to the order $\underline{\gamma}$ and the derivatives of order $\underline{\gamma}$ are Hölder continuous of the order $\gamma - \underline{\gamma}$.) We typically have that if a Gaussian process on $[0,1]^d$ is α-regular in this sense, then its RKHS unit ball $\mathbb{H}_1$ is contained in a Sobolev-type ball of regularity $\alpha + d/2$ (see for instance the concrete examples in the next subsection). If this is the case, then for every $\gamma \in [0, \alpha)$ the space $\mathbb{H}_1$ typically satisfies an entropy bound of the form (see, e.g., Edmunds and Triebel (1996))

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|_{C^\gamma}) \leq K_\gamma \varepsilon^{-\frac{2d}{d+2(\alpha-\gamma)}} \tag{3.3}$$

for some $K_\gamma > 0$. Here $\|\cdot\|_{C^\gamma}$ denotes the usual Hölder norm on $C^\gamma[0,1]^d$ (see e.g. Van der Vaart and Wellner (1996) for its precise definition).

Theorem 3.1. *Suppose that for $\alpha > d/2$ the process W takes values in $C^\gamma([0,1]^d)$ for every $\gamma < \alpha$ and its RKHS unit ball $\mathbb{H}_1$ satisfies the entropy bound (3.3) for every $\gamma \in [0, \alpha)$.*[1] *If (3.2) holds for numbers $\varepsilon_n \to 0$ such that $n\varepsilon_n^4 \to 0$, then with Δ_n and I_{σ_0} given by (2.1) we have*

$$\sup_B \left| \Pi(\sqrt{n}(\sigma - \sigma_0) \in B, f \in \mathcal{F} | Y_1, \ldots, Y_n) - N(\Delta_n, I_{\sigma_0}^{-1})(B) \right| \xrightarrow{P_0} 0$$

and $n \to \infty$, where the supremum is taken over all measurable subsets $B \subset \mathbb{R}$.

Proof. We first remark that if (3.2) holds for the sequence ε_n then it also holds for larger sequences, in particular for $\varepsilon_n' = \varepsilon_n \vee n^{-\frac{\alpha}{d+2\alpha}}$. Since $\alpha > d/2$, this new

[1] The proof shows that in fact it is sufficient if there exist $\alpha > \gamma > d/2$ such that W takes values in $C^\gamma([0,1]^d)$, and (3.3) holds for that α and γ and for α and $\gamma = 0$.

sequence satisfies $n(\varepsilon_n')^4 \to 0$ as well. Therefore, we can assume without loss of generality that $\varepsilon_n \geq n^{-\frac{\alpha}{d+2\alpha}}$ in the remainder of the proof.

We apply Theorem 2.2. It is well known that (3.2) implies that condition (1.2) is fulfilled with $\tilde{\varepsilon}_n = \varepsilon_n$ (see Van der Vaart and Van Zanten (2008b), Lemma 5.3). To prove that there exists sieves $\mathcal{F}_n$ such that (1.3)–(1.5) are satisfied we exploit the fact that by assumption we can view W as a Gaussian random element in the Banach space $(C^\gamma[0,1]^d, \|\cdot\|_{C^\gamma})$ for $\gamma < \alpha$. Since $C[0,1]^d$ is the completion of $C^\gamma[0,1]^d$ with respect to the $\|\cdot\|_\infty$-norm and $\|\cdot\|_\infty \leq \|\cdot\|_{C^\gamma}$, we have that the RKHS of W viewed as a $C^\gamma[0,1]^d$-valued Gaussian random element coincides with the RKHS $\mathbb{H}$ of W viewed as continuous Gaussian process. This follows from Lemma 8.1 in Van der Vaart and Van Zanten (2008b).

Since $\alpha > d/2$ by assumption, there exists a γ such that $\alpha > \gamma > d/2$. Now set $\delta_n = n^{-\frac{\alpha-\gamma}{d+2\alpha}}$ and $\mathcal{F}_n = M\sqrt{n}\varepsilon_n\mathbb{H}_1 + \delta_n C_1^\gamma$, where M is a constant to be determined below and C_1^γ is the unit ball in $C^\gamma[0,1]^d$. We claim that if M is chosen large enough, then conditions (1.3)–(1.5) hold true.

By the relation between the entropy of the RKHS unit ball and small ball probabilities established by Li and Linde (1999), assumption (3.3) implies that $\mathbb{P}(\|W\|_{C^\gamma} < \delta) \geq \exp(-D\delta^{-d/(\alpha-\gamma)})$ for some $D > 0$. It follows that

$$-\log \mathbb{P}(\|W\|_{C^\gamma} < \delta_n) \lesssim \delta_n^{-\frac{d}{\alpha-\gamma}} = n^{\frac{d}{d+2\alpha}} \leq n\varepsilon_n^2.$$

Hence, by the Borell-Sudakov inequality (see Van der Vaart and Van Zanten (2008b)) and the fact that for the standard normal distribution function Φ we have $\Phi^{-1}(y) \geq -\sqrt{(5/2)\log(1/y)}$ for small y, we have that condition (1.3) is fulfilled with $\tilde{\varepsilon}_n = \varepsilon_n$, provided M is chosen large enough.

For the entropy conditions we note that by assumption (3.3) (applied with $\gamma = 0$ this time) and known entropy bounds for Hölder balls (see for instance Van der Vaart and Wellner (1996)), we have

$$\begin{aligned}\log(2\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) &\leq \log N(\varepsilon, M\sqrt{n}\varepsilon_n\mathbb{H}_1, \|\cdot\|_\infty) + \log N(\varepsilon, \delta_n C_1^\gamma, \|\cdot\|_\infty) \\ &\lesssim \Big(\frac{\sqrt{n}\varepsilon_n}{\varepsilon}\Big)^{\frac{2d}{d+2\alpha}} + \Big(\frac{\delta_n}{\varepsilon}\Big)^{\frac{d}{\gamma}}.\end{aligned}$$

The right-hand side with ε_n substituted for ε is bounded by a constant times $n^{d/(d+2\alpha)} + (\delta_n/\varepsilon_n)^{d/\gamma}$. Both terms in this sum are bounded by $n\varepsilon_n^2$ by the lower bound assumption on ε_n and the definition of δ_n. Hence, condition (1.4) holds. The inequality in the last display also shows that for $a > 0$,

$$\int_0^{a\varepsilon_n} \sqrt{\log(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty)}\, d\varepsilon \lesssim n^{\frac{d}{2d+4\alpha}}\varepsilon_n + \delta_n^{\frac{d}{2\gamma}}\varepsilon_n^{\frac{2\gamma-d}{2\gamma}}.$$

Since $\alpha \geq d/2$ and $n\varepsilon_n^4 \to 0$, the first term on the right converges to 0. Since $\gamma > d/2$, the second term vanishes as well. This covers condition (1.5). □

3.2. Specific Gaussian priors

In this subsection we verify the conditions of Theorem 3.1 for two particular examples of Gaussian process priors on f. In the first example we investigate a

Matérn prior on a multivariate regression function. In the second example we consider the case $d = 1$ and choose a Riemann-Liouville type prior.

3.2.1. Matérn prior

The Matérn process $(W_t : t \in [0,1]^d)$ with parameter $\alpha > 0$ is the zero-mean, stationary Gaussian process with covariance function

$$\mathbb{E}W(x)W(y) = \int_{\mathbb{R}^d} e^{i\lambda^T(x-y)}\,\mu(d\lambda),$$

where the spectral measure μ is given by

$$\mu(\lambda) = \frac{d\lambda}{(1 + \|\lambda\|^2)^{\alpha + d/2}}.$$

A special case is the Ornstein-Uhlenbeck process, which is the case $d = 1$, $\alpha = 1/2$. The Matérn process is a popular prior in Bayesian nonparametrics, see for instance Rasmussen and Williams (2006) and the references therein.

It is not difficult to see that there exists a version of the Matérn process with parameter $\alpha > 0$ that takes its values in $C^\gamma([0,1]^d)$ for any $\gamma < \alpha$, see Van der Vaart and Van Zanten (2011). The RKHS unit ball of the Matérn process is included in a Sobolev ball of regularity $\alpha + d/2$, cf. Section 4.3 of Van der Vaart and Van Zanten (2011). For $\gamma < \alpha$, the metric entropy relative to the C^γ-norm of such a Sobolev ball satisfies (3.3) (see Theorem 3.3.2 on p. 105 in Edmunds and Triebel (1996)).

Now suppose that for $\beta > 0$, the true regression function is β-regular both in Hölder and Sobolev sense, i.e. $f_0 \in C^\beta([0,1]^d) \cap H^\beta([0,1]^d)$. The Hölder space was defined above and the Sobolev space $H^\beta([0,1]^d)$ consists of all functions f on $[0,1]^d$ that can be extended to a function f on all of $\mathbb{R}^d$ with Fourier transform $\hat{f}$ satisfying

$$\int |\hat{f}(\lambda)|^2 (1 + \|\lambda\|^2)^\beta \, d\lambda < \infty.$$

It is shown in Section IV of Van der Vaart and Van Zanten (2011) that for such f_0 the inequality (3.2) holds for ε_n proportional to $n^{-(\alpha\wedge\beta)/(d+2\alpha)}$.

It is easily verified that in this situation the conditions of Theorem 3.1 are satisfied if the regularity α of the prior and the regularity β of the true regression function satisfy the conditions

$$\begin{aligned} \frac{\alpha}{d} &> \frac{1}{2}, \\ \frac{\beta}{d} &> \frac{\alpha}{2d} + \frac{1}{4}, \end{aligned} \tag{3.4}$$

and hence the BvM statement for the marginal posterior distribution of σ holds under these conditions.

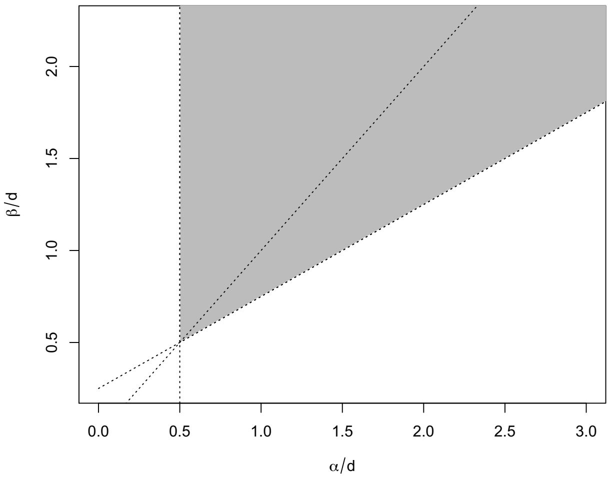

Fig 1. *The shaded area describes the values for the smoothness β of the true regression function f_0 and the regularity α of the Gaussian prior for which we have shown the BvM result holds.*

The collection of α's and β's satisfying (3.4) is sketched in Figure 1. The figure makes clear that for the BvM result to hold, it is not necessary to estimate the regression function f_0 at an optimal rate. In particular, it is not necessary that the smoothness α of the prior matches the smoothness β of the unknown regression function exactly. An arbitrary amount of undersmoothing ($\beta > \alpha$) is allowed and also some degree of oversmoothing ($\beta < \alpha$).

We note that it is not ruled out that the area for which BvM holds is actually larger than what we found. Using our general theorems it does not seem possible however to shed more light on this issue. Possibly more insight can be obtained by a more detailed analysis, tailored to the particular statistical problem and prior, in the spirit of Castillo (2012b).

3.2.2. Riemann-Liouville prior

In this subsection we consider the case $d = 1$, i.e. the true regression function is an unknown element $f_0 \in C[0, 1]$.

For $\alpha > 0$ and W a standard Brownian motion, the Riemann-Liouville process with parameter α is defined by

$$R_t^\alpha = \int_0^t (t-s)^{\alpha-1/2}\, dW_s.$$

It can be interpreted as the $(\alpha - 1/2)$-fold iterated integral of Brownian motion. The use of such priors is well established and goes back at least to Wahba (1978).

The process R^α and its higher derivatives (if they exist) vanish at zero. In order to enlarge the class of functions that are well approximated by the process we modify it slightly, following Van der Vaart and Van Zanten (2008a). Let $\underline{\alpha}$ be the biggest integer strictly smaller than α, and let $Z_1, \dots, Z_{\underline{\alpha}+1}$ be independent standard normal random variables, independent of the Riemann-Liouville process R^α. Define the Riemann-Liouville-type process X as follows:

$$X_t = \sum_{k=0}^{\underline{\alpha}+1} Z_k t^k + R_t^\alpha.$$

The process $(X_t : t \in [0,1])$ is zero-mean Gaussian and can be seen as a random element in $C[0,1]$.

Since Brownian motion has "regularity" $1/2$ the Riemann-Liouville process with parameter α is expected to be "regular" of order α in an appropriate sense. Indeed it can be shown that the process R^α, and hence also the process X, has a version that take values in $C^\gamma[0,1]$ for all $\gamma < \alpha$, cf. Lifshits and Simon (2005). The RKHS unit ball of X is a Sobolev-type ball of regularity $\alpha + 1/2$, cf. e.g. Van der Vaart and Van Zanten (2008a), and hence satisfies (3.3) with $d = 1$. Alternatively, the entropy bound (3.3) follows from the bound on the small ball probability of the Riemann-Liouville process with respect to the C^γ-norm given by Lifshits and Simon (2005) in combination with the result of Li and Linde (1999).

Upper bounds for the left hand side of (3.2) in this case are given in Van der Vaart and Van Zanten (2008a) and Castillo (2012a). If f_0 is in $C^\beta[0,1]$ for some $\beta \geq \alpha$, then the left hand side of (3.2) is bounded from above by a multiple of $\varepsilon_n^{-1/\alpha}$. For $\beta < \alpha$, the upper bound in Castillo (2012a) is $\varepsilon_n^{-(2\alpha-2\beta+1)/\beta} \log(1/\varepsilon_n)$. It follows that condition (3.2) is satisfied for ε_n a multiple of $(\log n/n)^{\beta/(1+2\alpha)}$ if $\beta < \alpha$ and for ε_n a multiple of $n^{-\alpha/(1+2\alpha)}$ if $\beta \geq \alpha$. These conditions are almost the same as in the Matérn prior case. The log factor does not affect the pairs (α, β) for which the inequalities are true. We thus obtain that for the Riemann-Liouville prior as well, the BvM statement of Theorem 3.1 holds if the regularity β of the truth and the regularity α of the prior as related as in (3.4), for $d = 1$. Again, Figure 1 visualizes the set of α's and β's.

4. Hierarchical spline-based priors

We consider again the case $\mathcal{X} = [0,1]^d$ in this section and investigate a spline prior on f. Such priors were considered for nonparametric regression for instance by Huang (2004) and De Jonge and Van Zanten (2012), where it was shown that when properly constructed, they can yield adaptive, rate-optimal procedures for estimating the regression function. Here we show that it is possible to simultaneously have BvM for the error standard deviation.

We fix an order $q \geq 2$ and for $m \in \mathbb{N}$, consider the space S_m of polynomial splines of order q with simple knots at the points $1/m, 2/m, \dots, (m-1)/m$. A function $s : [0,1] \to \mathbb{R}$ belongs to S_m if there exist polynomials $p_1, \dots, p_m$ of degree at most $q-1$ such that $s(x) = p_j(x)$ for $x \in [(j-1)/m, jm)$ and s is $q-2$ times continuously differentiable. The space S_m has dimension $J_m = q + m - 1$, cf. Theorem 4.4 of Schumaker (1981). A convenient basis of the space is given by the so-called B-splines. The exact definition of these functions (see Theorem 4.9 of Schumaker (1981)) is not of importance to us here. Important properties of B-splines are that they are nonnegative and supported on relative small parts of the domain and that the sum of all B-splines at any given location equals one. More precisely, they form a partition of unity: if we denote the B-splines by $B_1^m, \dots, B_{J_m}^m$, then $\sum_{j=1}^{J_m} B_j^m(x) = 1$ for all $x \in [0,1]$. As a consequence, the supremum norm $\|s\|_\infty$ of a function $s \in S_m$ of the form $s = \sum c_j B_j^m$ is bounded by the supremum norm of its B-spline coefficients $\|c\|_\infty = \max |c_j|$.

Functions of several variables can be dealt with using tensor product splines. For $d \geq 2$ we define the tensor product space $\mathcal{S}_m = S_m \otimes \cdots \otimes S_m$ (d times), with S_m the space of univariate splines defined above. The space $\mathcal{S}_m$ has dimension J_m^d and a basis is given by the tensor-product B-splines

$$B_j^m(x_1, \dots, x_d) = B_{j_1}^m(x_1) \cdots B_{j_d}^m(x_d), \quad 1 \leq j_i \leq J_m.$$

Slightly abusing notation these multivariate B-splines are denoted by $B_1^m, \dots, B_{J_m^d}^m$. It is easy to see that we again have the partition of unity property and hence also for $d \geq 2$ it holds that the supremum norm of a function in $\mathcal{S}_m$ is bounded by the supremum norm of its B-spline coefficients.

We define the prior Π_f on f as the law of the random spline process W defined by

$$W(x) = \sum_{j=1}^{J_M^d} \xi_j B_j^M(x), \quad x \in [0,1]^d,$$

where $\xi_1, \xi_2, \dots$ are independent, standard normal random variables and M^d is a geometric variable, independent of the ξ_j's. Theorem 4.2 of De Jonge and Van Zanten (2012) asserts that if $f_0 \in C^\beta([0,1]^d)$ for some $\beta \leq q$, then corresponding posterior distribution satisfies (1.1) for ε_n equal to $n^{-\beta/(d+2\beta)}$, up to a logarithmic factor. In particular, with this prior we achieve nearly rate-optimal, adaptive estimation of the regression function for regularities up to the order of the splines that are used. We can now prove that if the regularity of the regression function is larger than the dimension of the design space, we simultaneously have BvM for σ.

Theorem 4.1. *Suppose that $f_0 \in C^\beta([0,1]^d)$ for some $\beta \in (d, q]$. Then with Δ_n and I_{σ_0} given by (2.1) we have*

$$\sup_B \left| \Pi(\sqrt{n}(\sigma - \sigma_0) \in B, f \in \mathcal{F} | Y_1, \dots, Y_n) - N(\Delta_n, I_{\sigma_0}^{-1})(B) \right| \overset{P_0}{\to} 0$$

and $n \to \infty$, where the supremum is taken over all measurable subsets $B \subset \mathbb{R}$.

Proof. It was proved in De Jonge and Van Zanten (2012) (see Theorem 4.2 in that paper) that if $f_0 \in C^\beta([0,1]^d)$ for $\beta \leq q$, then for sequences $\tilde{\varepsilon}_n$ and $\bar{\varepsilon}_n$ that are both up to a logarithmic factor equal to $n^{-\beta/(d+2\beta)}$, it holds that for every

$C > 1$ there exists a constant $D > 0$ and sets $U_n \subset C[0,1]$ such that

$$\begin{aligned}\mathbb{P}(\|W - f_0\|_\infty \le 2\tilde{\varepsilon}_n) &\ge \exp(-n\tilde{\varepsilon}_n^2),\\ \mathbb{P}(W \notin U_n) &\le \exp(-Cn\tilde{\varepsilon}_n^2),\\ \log N(2\bar{\varepsilon}_n, U_n, \|\cdot\|_\infty) &\le Dn\bar{\varepsilon}_n^2.\end{aligned}$$

So we see that conditions (1.2)–(1.4) of Theorem 2.2 are satisfied. The sets U_n are certain unions of enlarged RKHS balls corresponding to the Gaussian process that is obtained by conditioning the process W on the gridsize variable M. Inspection of the proof of Theorem 4.2 of De Jonge and Van Zanten (2012) however shows that condition (1.5) does not hold for the U_n.

Fix $C > 1$. To construct new, slightly smaller sieves we take constants $K, L > 0$, determined further below, and define

$$V_n = \bigcup_{m \le (Kn\tilde{\varepsilon}_n^2)^{1/d}} V_n^m, \quad V_n^m = \Big\{ \sum_{j \le J_m^d} c_j B_j^m : \max |c_j| \le L\sqrt{n}\tilde{\varepsilon}_n \Big\}.$$

Then we set $\mathcal{F}_n = U_n \cap V_n$. We claim that conditions (1.3)–(1.5) are satisfied for these sets.

We have $\Pi(\mathcal{F}_n^c) \le \mathbb{P}(W \notin U_n) + \mathbb{P}(W \notin V_n)$. The first probability is bounded by $\exp(-Cn\tilde{\varepsilon}_n^2)$ and by construction we have

$$\mathbb{P}(W \notin V_n) \le \mathbb{P}(M > (Kn\tilde{\varepsilon}_n^2)^{1/d}) + \sum_{m \le (Kn\tilde{\varepsilon}_n^2)^{1/d}} \mathbb{P}\Big(\max_{j \le J_m^d} |Z_j| > L\sqrt{n}\tilde{\varepsilon}_n \Big).$$

Hence, since the variable M^d is geometric and $\mathbb{P}(\max_{j \le J_m^d} |Z_j| > L\sqrt{n}\tilde{\varepsilon}_n) \lesssim m^d \exp(-L^2 n\tilde{\varepsilon}_n^2/2)$,

$$\mathbb{P}(W \notin V_n) \lesssim e^{-cKn\tilde{\varepsilon}_n^2} + (Kn\tilde{\varepsilon}_n^2)^{1+1/d} e^{-\frac{1}{2}Ln\tilde{\varepsilon}_n^2}$$

for some $c > 0$. For K, L large enough this is bounded by $\exp(-Cn\tilde{\varepsilon}_n^2)$ as well, and it follows that condition (1.3) is fulfilled.

It is clear that the sieves $\mathcal{F}_n$ satisfy condition (1.4), since the are contained in the U_n. Next, observe that for $\delta > 0$,

$$N(\delta, V_n, \|\cdot\|_\infty) \le \sum_{m \le (Kn\tilde{\varepsilon}_n^2)^{1/d}} N(\delta, V_n^m, \|\cdot\|_\infty).$$

Since the supremum norm of a spline in $\mathcal{S}_m$ is bounded by the supremum norm of its B-spline coefficients,

$$N(\delta, V_n^m, \|\cdot\|_\infty) \le (N(\delta, [-L\sqrt{n}\tilde{\varepsilon}_n, L\sqrt{n}\tilde{\varepsilon}_n], |\cdot|))^{J_m^d} \le \Big(\frac{2L\sqrt{n}\tilde{\varepsilon}_n}{\delta} \Big)^{J_m^d}.$$

It follows that for every $a, \varepsilon > 0$,

$$\int_0^{a\varepsilon} \sqrt{\log N(\delta, V_n, \|\cdot\|_\infty)}\, d\delta \lesssim a\varepsilon \log n + n\tilde{\varepsilon}_n^2 \int_0^{a\varepsilon} \sqrt{\log\Big(\frac{2L\sqrt{n}\tilde{\varepsilon}_n}{\delta} \Big)}\, d\delta.$$

It is easily checked that the integral on the right is bounded by a constant times $a\varepsilon \log(2L\sqrt{n}\tilde{\varepsilon}_n/(a\varepsilon))$. All together we find that for $\varepsilon_n = \tilde{\varepsilon}_n \vee \bar{\varepsilon}_n$,

$$\int_0^{a\varepsilon_n} \sqrt{\log N(\delta, V_n, \|\cdot\|_\infty)}\, d\delta \lesssim \varepsilon_n^3 n \log n.$$

Since $\varepsilon_n \lesssim n^{-\beta/(d+2\beta)} \log^p n$ for some $p > 0$, the right-hand side converges to 0 if $\beta > d$. This covers condition (1.5) and also shows that $n\varepsilon_n^4 \to 0$, as required. □

We remark that the condition $\beta > d$ is used for technical reasons in the proof, to control the last entropy integral appearing in the proof. This does not rule out the possibility that the statement of the theorem is true for a larger range of β's.

5. Proof of the general theorem

In this section we give the proof of Theorem 2.1.

It is convenient to describe the model by the parameter (θ, f) with $\theta = 1/\sigma^2$. For this parametrization the log-likelihood is given by

$$\ell_n(\theta, f) = \frac{n}{2} \log \frac{\theta}{2\pi} - \frac{\theta}{2} \sum_{i=1}^n (Y_i - f(x_i))^2.$$

The first step in the proof is finding an appropriate expansion for the log-likelihood ratio $\Lambda_n(\theta, f) = \ell_n(\theta, f) - \ell_n(\theta_0, f_0)$. We define an inner product $\langle \cdot, \cdot \rangle_L$ on pairs (θ, f) of inverse variances and regression functions by

$$\langle (\theta, f), (\psi, g) \rangle_L = \frac{\theta\psi}{2\theta_0^2} + \frac{\theta_0}{n} \sum_{i=1}^n f(x_i) g(x_i).$$

The corresponding norm is denoted by $\|\cdot\|_L$, so

$$\|\theta, f\|_L^2 = \frac{\theta^2}{2\theta_0^2} + \theta_0 \|f\|_n^2.$$

Note that although it is not made explicit in the notation, the inner product and the norm depend on the sample size n (and on the true parameter θ_0).

Straightforward algebra yields the following lemma.

Lemma 5.1. *We have*

$$\Lambda_n(\theta, f) = -\frac{n}{2} \|\theta - \theta_0, f - f_0\|_L^2 + \sqrt{n} W_n(\theta - \theta_0, f - f_0) + R_n(\theta, f),$$

where

$$W_n(\theta, f) = -\frac{\theta}{2\theta_0\sqrt{n}} \sum_{i=1}^n (Z_i^2 - 1) + \sqrt{\frac{\theta_0}{n}} \sum_{i=1}^n f(x_i) Z_i$$

and

$$R_n(\theta, f) = \frac{n}{2}\Big(\log\theta - \log\theta_0 - \frac{\theta-\theta_0}{\theta_0} + \frac{(\theta-\theta_0)^2}{2\theta_0^2}\Big)$$
$$- \frac{1}{2}n(\theta-\theta_0)\|f-f_0\|_n^2 + \frac{\theta-\theta_0}{\sqrt{\theta_0}}\sum_{i=1}^n (f(x_i)-f_0(x_i))Z_i.$$

We are now in the situation that we can apply Theorem 1 of Castillo (2012a). Strictly speaking this theorem does not allow the dependence of the inner product $\langle\cdot,\cdot\rangle_L$ on n that we have, but inspection of Castillo's proof shows that this causes no problems. Since our LAN-norm has the property that the norm $\theta \mapsto \|\theta, 0\|_L$ on $\mathbb{R}$ is independent of n, only minor adaptations of that proof are necessary. We note that our change of variables $\theta = 1/\sigma^2$ helps to establish a direct connection with the setup of Castillo (2012a), since the map W_n defined in Lemma 5.1 is linear in θ.

Castillo's theorem asserts that if there exists positive numbers $\delta_n \to 0$ such that $n\delta_n^2 \to \infty$ and measurable subsets $\mathcal{F}_n \subset \mathcal{F}$ such that

$$\Pi((\theta, f) \in (0,\infty)\times\mathcal{F}_n : \|\theta-\theta_0, f-f_0\|_L \le \delta_n \,|\, Y_1,\dots,Y_n) \overset{P_0}{\to} 1, \tag{5.1}$$

$$\Pi^{\theta=\theta_0}(f\in\mathcal{F}_n : \|0, f-f_0\|_L \le \delta_n/\sqrt{2} \,|\, Y_1,\dots,Y_n) \overset{P_0}{\to} 1, \tag{5.2}$$

$$\sup_{\substack{(\theta,f)\in(0,\infty)\times\mathcal{F}_n:\\ \|\theta-\theta_0, f-f_0\|_L\le\delta_n}} \frac{|R_n(\theta,f)-R_n(\theta_0,f)|}{1+n(\theta-\theta_0)^2} \overset{P_0}{\to} 0, \tag{5.3}$$

then

$$\sup_B \Big|\Pi(\sqrt{n}(\theta-\theta_0)\in B, f\in\mathcal{F}|Y_1,\dots,Y_n) - N\Big(\frac{W_n(1,0)}{\|1,0\|_L^2}, \frac{1}{\|1,0\|_L^2}\Big)(B)\Big| \overset{P_0}{\to} 0. \tag{5.4}$$

The next step is to show that conditions (5.1)–(5.3) hold for δ_n equal to a constant times ε_n under the assumptions of Theorem 2.1.

Since $\sqrt{x+y} \le \sqrt{x}+\sqrt{y}$ for $x, y \ge 0$, we have $\|\theta-\theta_0, f-f_0\|_L \le C(|\theta-\theta_0| + \|f\|_n)$, for a constant $C > 0$ only depending on θ_0. It follows that under assumptions (2.3) and (2.4), conditions (5.1) and (5.2) hold for δ_n a multiple of ε_n.

Next we consider (5.3). Define $V_n = \{(\theta,f)\in(0,\infty)\times\mathcal{F}_n : \|\theta-\theta_0, f-f_0\|_L \le \delta_n\}$. We consider the three terms in the definition of R_n in the statement of Lemma 5.1 separately. For $\theta_0 \in V_n$ it holds that $|\theta-\theta_0|$ is bounded by a multiple of δ_n. By Taylor's formula, the first term in the definition of R_n is $nO(|\theta-\theta_0|^3)$ for θ close to θ_0, and hence the first term is bounded by a multiple of $(1+n(\theta-\theta_0)^2)\delta_n$ on V_n. For the second term, note that $x \mapsto x/(1+nx^2)$ is maximal at $x = n^{-1/2}$, and equal to $n^{-1/2}/2$ at that point. It follows that

$$\sup_{(\theta,f)\in V_n} \frac{n|\theta-\theta_0|\|f-f_0\|_n^2}{1+n(\theta-\theta_0)^2} \le \frac{1}{2}\sqrt{n}\sup_{(\theta,f)\in V_n}\|f-f_0\|_n^2 \le \frac{\sqrt{n}\delta_n^2}{2\theta_0}.$$

Similarly, the supremum over V_n of third term divided by $1 + n(\theta - \theta_0)^2$ is bounded by

$$\frac{1}{2\sqrt{\theta_0}} \sup_{\substack{f \in \mathcal{F}_n \\ \sqrt{\theta}_0 \| f - f_0 \|_n \leq \delta_n}} |\mathbb{G}_n f - \mathbb{G}_n f_0|,$$

where $\mathbb{G}_n$ is the Gaussian random map defined by

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(x_i) Z_i.$$

The norm $\|\cdot\|_n$ is precisely the natural semi-norm associated with the Gaussian process $\mathbb{G}_n$, in the sense that $\mathbb{E}_0(\mathbb{G}_n f - \mathbb{G}_n g)^2 = \|f - g\|_n^2$. Therefore, the well-known maximal inequality for sub-Gaussian processes, cf. e.g. Van der Vaart and Wellner (1996), Corollary 2.2.8, implies that

$$\mathbb{E}_0 \sup_{\substack{f \in \mathcal{F}_n \\ \sqrt{\theta}_0 \| f - f_0 \|_n \leq \delta_n}} |\mathbb{G}_n f - \mathbb{G}_n f_0| \leq K \int_0^{\delta_n / \sqrt{\theta}_0} \sqrt{\log N(\delta, \mathcal{F}_n, \|\cdot\|_n)}\, d\delta$$

for some constant $K > 0$. All together we conclude that the left-hand side of (5.3) is

$$O_{\mathbb{P}_0}\Big(\delta_n + \sqrt{n}\delta_n^2 + \int_0^{\delta_n / \sqrt{\theta}_0} \sqrt{\log N(\delta, \mathcal{F}_n, \|\cdot\|_n)}\, d\delta\Big)$$

for $n \to \infty$. For δ_n a multiple of ε_n this is $o_{\mathbb{P}_0}(1)$ under the assumptions of the theorem, hence (5.3) holds as well.

We have now established that (5.4) holds under the conditions of Theorem 2.1. Next, observe that $\|1, 0\|_L^2 = 1/(2\theta_0^2)$ and

$$\frac{W_n(1,0)}{\|1,0\|_L^2} = -\frac{\theta_0}{\sqrt{n}} \sum_{i=1}^n (Z_i^2 - 1) \Rightarrow N(0, 2\theta_0^2)$$

under $\mathbb{P}_0$, by the central limit theorem. The statement of the theorem then follows by an application of the lemma below, which gives a total variation version of the delta method, tailored to our situation. We apply the lemma with X_n a random variable which has the posterior distribution of θ as law, $x_0 = \theta_0$, $\mu_n = W_n(1,0)/\|1,0\|_L^2$, $\sigma^2 = 1/\|1,0\|_L^2 = 2\theta_0^2$ and $f(x) = 1/\sqrt{x}$. The lemma deals with the total variation distance between deterministic distributions. We can use it in our stochastic setting since $W_n(1,0)/\|1,0\|_L^2$ converges in distribution and hence is uniformly tight.

We denote the total variation distance between two probability measure μ and ν by $d_{TV}(\mu, \nu)$ and the law, or distribution of a random variable X by $\mathcal{L}(X)$.

Lemma 5.2. *Let X_n be a sequence of random variables such that*

$$d_{TV}(\mathcal{L}(\sqrt{n}(X_n - x_0)), N(\mu_n, \sigma^2)) \to 0, \tag{5.5}$$

for $x_0 \in \mathbb{R}$, $\sigma^2 > 0$ and μ_n a bounded sequence. Let $f : \mathbb{R} \to \mathbb{R}$ be a function that is twice continuously differentiable on a neighborhood of x_0 and $f'(x_0) \neq 0$. Then

$$d_{TV}(\mathcal{L}(\sqrt{n}(f(X_n) - f(x_0))), N(f'(x_0)\mu_n, (\sigma f'(x_0))^2)) \to 0.$$

Proof. We suppose for definiteness that $f'(x_0) > 0$. It follows from the assumptions on f that there exist neighborhoods U and V of x_0 and $f(x_0)$ such that f is an invertible (in this case increasing) bijection between U and V. The distribution $N(x_0 + \mu_n/\sqrt{n}, \sigma^2/n$ concentrates around x_0 as $n \to \infty$. Hence, by (5.5), so does $\mathcal{L}(X_n)$ and hence the law $\mathcal{L}(f(X_n))$ concentrates around $f(x_0)$. Therefore, we only need to prove that

$$\sup_{B\subset V} |\mathbb{P}(f(X_n) \in B) - N(f(x_0) + \mu_n f'(x_0)/\sqrt{n}, (f'(x_0))^2\sigma^2/n)(B)| \to 0,$$

or, equivalently,

$$\sup_{A\subset U} |\mathbb{P}(X_n \in A) - N(f(x_0) + \mu_n f'(x_0)/\sqrt{n}, (f'(x_0))^2\sigma^2/n)(f(A))| \to 0.$$

Using (5.5), a change of variables and some straightforward algebra we then see that it suffices to show that

$$\int_U \Big|\frac{1}{\tau_n}\varphi\Big(\frac{f'(x_0)(x - x_0) - \delta_n}{\tau_n}\Big) f'(x_0) - \frac{1}{\tau_n}\varphi\Big(\frac{f(x) - f(x_0) - \delta_n}{\tau_n}\Big) f'(x)\Big|\, dx \to 0,$$

where φ denotes the standard normal density, $\delta_n = \mu_n f'(x_0)/\sqrt{n}$ and $\tau_n = \sigma f'(x_0)/\sqrt{n}$.

Consider the shrinking sets $U_n = \{x \in U : |x - x_0| \leq K_n\tau_n\}$ for a sequence $K_n \to \infty$ such that $K_n^3\tau_n \to 0$. For $x \in U_n^c$ it holds that $|f(x) - f(x_0)| \geq cK_n\tau_n$ for some $c > 0$ and hence

$$\int_{U_n^c} \frac{1}{\tau_n}\varphi\Big(\frac{f(x) - f(x_0) - \delta_n}{\tau_n}\Big) f'(x)\, dx \leq \int_{|z|>cK_n} \varphi(z - \mu_n/\sigma)\, dz \to 0.$$

Similarly,

$$\int_{U_n^c} \frac{1}{\tau_n}\varphi\Big(\frac{f'(x_0)(x - x_0) - \delta_n}{\tau_n}\Big)\, dx \to 0.$$

Since φ is Lipschitz and f is twice continuously differentiable we have

$$\frac{1}{\tau_n}\int_{U_n} \Big|\varphi\Big(\frac{f'(x_0)(x - x_0) - \delta_n}{\tau_n}\Big) - \varphi\Big(\frac{f(x) - f(x_0) - \delta_n}{\tau_n}\Big)\Big|\, dx \lesssim K_n^3\tau_n \to 0.$$

Finally, observe that by definition of U_n,

$$\begin{aligned}&\frac{1}{\tau_n}\int_{U_n} \varphi\Big(\frac{f(x) - f(x_0) - \delta_n}{\tau_n}\Big)|f'(x) - f'(x_0)|\, dx \\ &\quad\lesssim K_n \int_{U_n} \varphi\Big(\frac{f(x) - f(x_0) - \delta_n}{\tau_n}\Big)\, dx \lesssim K_n^2\tau_n \to 0.\end{aligned}$$

The proof is completed by combining the convergence statements derived in this paragraph. □

Acknowledgement

We are grateful to a referee for pointing out that our previous version of Theorem 3.1 could be significantly improved.

Appendix A: General contraction rate theorem for fixed design regression

A.1. Statement of the result

We consider the setting described in the first paragraph of Section 2.2. We now put a general prior Π_n on the pair (f,σ), not necessarily a product. Assume that for $0<a<b<\infty$, $\sigma_0 \in [a,b]$ and Π_n is concentrated on $[a,b]\times\mathcal{F}$. Let $\Pi_n(\cdot\,|\,Y_1,\dots,Y_n)$ be the corresponding posterior.

Theorem A.1. *Suppose we have sequences of positive numbers $\tilde\varepsilon_n, \bar\varepsilon_n \to 0$ such that $n(\tilde\varepsilon_n \wedge \bar\varepsilon_n)^2 \to \infty$. If for constants $c_1, c_2, c_3 > 0$ and sets $\mathcal{F}_n \subset \mathcal{F}$ we have*

$$\Pi_n\Big(\Big\|\frac{f-f_0}{\sigma_0}\Big\|_n \le \tilde\varepsilon_n, \Big|\frac{\sigma_0^2}{\sigma^2}-1\Big| \le \tilde\varepsilon_n\Big) \ge c_1 e^{-c_2 n\tilde\varepsilon_n^2},$$

$$\Pi_n(f\in\mathcal{F}_n^c, \sigma\in[a,b]) = o\Big(e^{-(c_2+7)n\tilde\varepsilon_n^2}\Big),$$

$$\log N(\bar\varepsilon_n, \mathcal{F}_n, \|\cdot\|_n) \le c_3 n\bar\varepsilon_n^2,$$

then for $\varepsilon_n = \tilde\varepsilon_n \vee \bar\varepsilon_n$ and every sufficiently large $M>0$,

$$\Pi_n\Big((f,\sigma)\in\mathcal{F}_n\times[a,b] : \Big\|\frac{f-f_0}{\sigma_0}\Big\|_n + \Big|\frac{\sigma^2}{\sigma_0^2}-1\Big| \le M\varepsilon_n \,|\, Y_1,\dots,Y_n\Big) \xrightarrow{P_0} 1.$$

as $n\to\infty$.

A.2. Proof of the theorem

We abbreviate $Y=(Y_1,\dots,Y_n)$, $f=(f(x_1),\dots,f(x_n))$, so that the likelihood is given by

$$p_{f,\sigma} = p_{f,\sigma}(Y) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\|Y-f\|^2},$$

and hence

$$\log\frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} = -\frac{n}{2}\log\frac{\sigma^2}{\sigma_0^2} - \frac12\Big(\frac{1}{\sigma^2}\|Y-f\|^2 - \frac{1}{\sigma_0^2}\|Y-f_0\|^2\Big). \tag{A.1}$$

Observe that for $M>0$, $\mathcal{F}_n\subset\mathcal{F}$ and $\varepsilon_n\to 0$, we have

$$\begin{aligned} 1 - \Pi_n\Big((f,\sigma)&\in\mathcal{F}_n\times[a,b] : \|f-f_0\|_n + |\sigma-\sigma_0| \le 2M\varepsilon_n \,|\, Y\Big) \\ &\le \Pi_n(f\in\mathcal{F}_n, \|f-f_0\|_n > M\varepsilon_n \,|\, Y) \\ &\quad + \Pi_n(f\in\mathcal{F}_n, |\sigma-\sigma_0| > M\varepsilon_n, \|f-f_0\|_n \le M\varepsilon_n \,|\, Y) \\ &\quad + \Pi_n(f\in\mathcal{F}\backslash\mathcal{F}_n \,|\, Y) =: I + II + III. \end{aligned} \tag{A.2}$$

We will show that these three terms vanish in $\mathbb{P}_0$-probability as $n \to \infty$ for M large enough.

In the following subsection we first lower bound the denominator in the expression for the posterior.

A.2.1. Lower bound for the denominator

For $B \subset \mathcal{F} \times [a, b]$, we can write

$$\Pi(B \,|\, Y_1, \ldots, Y_n) = \frac{\iint_B \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} \Pi(df, d\sigma)}{\iint \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} \Pi(df, d\sigma)}. \tag{A.3}$$

The following lemma deals with the denominator. In the proof, and also at other places below, we use the fact that for a standard Gaussian variable ξ and $a, b \in \mathbb{R}$, $b > -1$, we have

$$\mathbb{E} e^{a\xi - \frac{1}{2} b \xi^2} = \frac{1}{\sqrt{1+b}} e^{\frac{a^2}{2(1+b)}}. \tag{A.4}$$

Lemma A.2. *For $\varepsilon \in (0, 1/2)$,*

$$\mathbb{P}_0\Big(\iint \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} \, d\Pi \geq e^{-7n\varepsilon^2} \Pi\Big(\Big\| \frac{f - f_0}{\sigma_0} \Big\|_n \leq \varepsilon, \Big| \frac{\sigma_0^2}{\sigma^2} - 1 \Big| \leq \varepsilon \Big) \Big) \geq 1 - e^{-\frac{3}{4} n \varepsilon^2}.$$

Proof. let $\tilde{\Pi}$ a probability distribution on $\mathcal{F} \times [a, b]$ obtained by restricting Π to the set $\{(f, \sigma) : \|f - f_0\|_n \leq \sigma_0 \varepsilon, |\sigma_0^2/\sigma^2 - 1| \leq \varepsilon\}$ and renormalizing. The arithmetic-geometric mean inequality (or Jensen's inequality) implies that

$$\iint \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} \, d\tilde{\Pi} \geq \exp\Big(\iint \log \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} \, d\tilde{\Pi} \Big).$$

It follows that for $x > 0$,

$$\mathbb{P}_0\Big(\iint \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} \, d\tilde{\Pi} \leq e^{-x} \Big) \leq \mathbb{P}_0\Big(- \iint \log \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} \, d\tilde{\Pi} > x \Big).$$

We have (see (A.1)), with $h = (f - f_0)/\sigma_0$ and $Z = (Z_1, \ldots, Z_n)$,

$$-\log \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}} = \frac{n}{2} \log \frac{\sigma^2}{\sigma_0^2} + \frac{1}{2} \Big(\Big(\frac{\sigma_0^2}{\sigma^2} - 1 \Big) \|Z\|^2 + 2 \frac{\sigma_0^2}{\sigma^2} \langle Z, h \rangle + \frac{\sigma_0^2}{\sigma^2} \|h\|^2 \Big).$$

Hence, the last probability is bounded by $\mathbb{P}(\langle Z, v \rangle - (1/2) w \|Z\|^2 > y)$, for v the vector with coordinates

$$v_i = \iint \frac{\sigma_0^2}{\sigma^2} \Big(\frac{f(x_i) - f_0(x_i)}{\sigma_0} \Big) \, d\tilde{\Pi}$$

and

$$w = \iint \Big(1 - \frac{\sigma_0^2}{\sigma^2} \Big) \, d\tilde{\Pi},$$

$$y = x - n \iint \Big(\frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} + \frac{\sigma_0^2}{\sigma^2}\Big\|\frac{f-f_0}{\sigma_0}\Big\|_n^2\Big)\,d\tilde{\Pi}.$$

Note that it follows from the definition of $\tilde{\Pi}$ that $|w| \leq \varepsilon \leq 1/2$, hence $1+w \geq 1/2$, and $\|v\|^2 \leq (1+\varepsilon)^2 n\varepsilon^2 \leq 4n\varepsilon^2$. Therefore, by Markov's inequality, the probability is further bounded by

$$e^{-y}\prod \mathbb{E}e^{v_iZ_i - \frac{1}{2}wZ_i^2} = e^{-y}e^{\frac{\|v\|^2}{2(1+w)}}(1+w)^{-n/2} \leq e^{4n\varepsilon^2}e^{-y}(1+w)^{-n/2}.$$

Elementary manipulations show that $e^{-y}(1+w)^{-n/2}$ equals

$$\exp\Big(-x - \frac{n}{2}\Big(\iint\Big(\log\frac{\sigma_0^2}{\sigma^2} - \Big(\frac{\sigma_0^2}{\sigma^2} - 1\Big)\Big)\,d\tilde{\Pi} + \Big(\log(1+w) - w\Big)\Big) + n\iint\Big(\frac{\sigma_0^2}{\sigma^2}\Big\|\frac{f-f_0}{\sigma_0}\Big\|_n^2\Big)\,d\tilde{\Pi}\Big)$$

Since $0 \geq \log(1+x) - x \geq -x^2$ for $|x| \leq 1/2$, this is bounded by $\exp(-x + (9/4)n\varepsilon^2)$. It follows that

$$\mathbb{P}_0\Big(\iint \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}}\,d\tilde{\Pi} \leq e^{-x}\Big) \leq e^{-x+(25/4)n\varepsilon^2}.$$

The proof is completed by taking $x = 7n\varepsilon^2$ and recalling the definition of $\tilde{\Pi}$. □

The lemma implies that under the prior mass assumption of the theorem, it holds that

$$\iint \frac{p_{f,\sigma}}{p_{f_0,\sigma_0}}\,d\Pi_n \geq c_1 e^{-(c_2+7)n\tilde{\varepsilon}_n^2}$$

on an event A_n such that $\mathbb{P}_0(A_n) \to 1$.

We now proceed to prove that the terms on the right of (A.2) vanish.

A.2.2. Term I

In view of the preceding section it suffices to show that $\mathbb{E}_0\Pi_n(f \in \mathcal{F}_n, \|(f-f_0)/\sigma_0\|_n > M\varepsilon_n \,|\, Y)1_{A_n} \to 0$. For arbitrary tests φ_n the expectation is bounded by

$$\mathbb{E}_0\varphi_n + \frac{1}{c_1}e^{(c_2+7)n\varepsilon_n^2}\iint_{f\in\mathcal{F}_n, \|(f-f_0)/\sigma_0\|_n > M\varepsilon_n} \mathbb{E}_{f,\sigma}(1-\varphi_n)\,d\Pi_n.$$

The following lemma asserts we can construct tests for which both terms converge to 0.

Lemma A.3. *There exist tests φ_n such that $\mathbb{E}_0\varphi_n \to 0$ and*

$$e^{(c_2+7)n\varepsilon_n^2}\iint_{f\in\mathcal{F}_n, \|(f-f_0)/\sigma_0\|_n > M\varepsilon_n} \mathbb{E}_{f,\sigma}(1-\varphi_n)\,d\Pi_n \to 0$$

as $n \to \infty$.

Proof. For $f_1 \in \mathcal{F}$, let φ^{f_1} be the likelihood ratio test for testing the null (f_0, σ_0) against the alternative (f_1, σ_0), i.e. $\varphi^{f_1} = 1_{\|Y-f_1\|<\|Y-f_0\|}$. Then by the Gaussian tail bound,

$$\mathbb{E}_0\varphi^{f_1} = \mathbb{P}_0(2\langle Y-f_0, f_1-f_0\rangle > \|f_1-f_0\|^2) \le e^{-\frac{1}{8}n\left\|\frac{f_1-f_0}{\sigma_0}\right\|_n^2}.$$

On the other hand, straightforward computations show that for all $\sigma > 0$ and f such that $\|f-f_1\|_n \le \|f-f_0\|_n$,

$$\mathbb{E}_{f,\sigma}(1-\varphi^{f_1}) \le e^{-\frac{1}{8}n\frac{\sigma_0^2}{\sigma^2}\frac{\left(\left\|\frac{f-f_0}{\sigma_0}\right\|_n^2-\left\|\frac{f-f_1}{\sigma_0}\right\|_n^2\right)^2}{\left\|\frac{f_1-f_0}{\sigma_0}\right\|_n^2}}.$$

Now define the set $B = \{(f,\sigma) \in \mathcal{F}_n : \|(f-f_0)/\sigma_0\|_n \ge M\varepsilon_n\}$ and write $B = \bigcup_{i\ge M} S_i$, where $S_i = \{f \in \mathcal{F}_n : i\varepsilon_n \le \|(f-f_0)/\sigma_0\|_n < (i+1)\varepsilon_n\}$. For $i \ge M$, the entropy condition and the fact that $\varepsilon_n \ge \bar{\varepsilon}_n$ imply that S_i can be covered with, say, $N_i \le e^{c_4 n\varepsilon_n}$ $\|\cdot/\sigma_0\|_n$-balls of radius ε_n. Let S_i be the collection of the N_i center points of the balls. Let $\tau_i : S_i \to C_i$ be a map which maps a point in S_i to a closest point in C_i. Note that by construction $\|(f-\tau_i(f))/\sigma_0\|_n \le \varepsilon_n$ for every $f \in S_i$. Define the sequence of tests $\varphi_n = \sup_{i\ge M}\sup_{f\in C_i}\varphi^f$.

We have

$$\mathbb{E}_0\varphi_n \le \sum_{i\ge M}\sum_{f\in C_i}\mathbb{E}_0\varphi^f \le \sum_{i\ge M} e^{-(\frac{i^2}{8}-c_4)n\varepsilon_n^2}.$$

For M large enough this vanishes as $n \to \infty$. We also have, for $f \in S_i$ and $\sigma > 0$,

$$\mathbb{E}_{f,\sigma}(1-\varphi_n) \le \mathbb{E}_{f,\sigma}(1-\varphi^{\tau_i(f)}) \le e^{-\frac{1}{8}n\frac{\sigma_0^2}{\sigma^2}(i-1)^2\varepsilon_n^2}.$$

For $M > 0$ large enough we have $\mathbb{E}_0\varphi_n \to 0$. Also,

$$\begin{aligned}
&e^{(c_2+7)n\varepsilon_n^2}\iint_{f\in\mathcal{F}_n,\|(f-f_0)/\sigma_0\|_n>M\varepsilon_n}\mathbb{E}_{f,\sigma}(1-\varphi_n)\,d\Pi_n\\
&\quad= e^{(c_2+7)n\varepsilon_n^2}\sum_{i\ge M}\sup_{f\in S_i,\sigma\le b}\mathbb{E}_{f,\sigma}(1-\varphi_n)\\
&\quad\le e^{(c_2+7)n\varepsilon_n^2}\sum_{i\ge M-1} e^{-\frac{\sigma_0^2}{b^2}i^2n\varepsilon_n^2}
\end{aligned}$$

If M is large enough this vanishes as well if $n \to \infty$. □

A.2.3. Term II

For $\sigma_1 > 0$, $\sigma_1 \ne \sigma_0$, let $\varphi_n^{\sigma_1}$ be the likelihood ratio test for testing the null (f_0, σ_0) against the alternative (f_0, σ_1), i.e.

$$\varphi_n^{\sigma_1} = 1_{-\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2}-1\right)\left\|\frac{Y-f_0}{\sigma_0}\right\|^2 > -\frac{n}{2}\log\frac{\sigma_0^2}{\sigma_1^2}}.$$

Lemma A.4. *Suppose that $\sigma_0, \sigma_1, \sigma \in [a, b]$. There exists constants $\kappa_1, \kappa_2 > 0$, depending only on a and b, such that*

$$\mathbb{E}_0 \varphi_n^{\sigma_1} \le e^{-\kappa_1 n(\sigma_0^2/\sigma_1^2 - 1)^2}$$

and for f such that $\|f - f_0\|_n \le \varepsilon \le 1$,

$$\mathbb{E}_{f,\sigma}(1 - \varphi_n^{\sigma_1}) \le e^{-\frac{n}{2}(1 - \frac{\sigma^2}{\sigma_0^2})(\frac{\sigma_0^2}{\sigma_1^2} - 1) + \kappa_2 n(\frac{\sigma_0^2}{\sigma_1^2} - 1)^2 + n(\frac{\sigma_0^2}{\sigma_1^2} - 1)\varepsilon^2}.$$

Proof. For $\lambda \in (0, 1)$ we have, by Markov's inequality and (A.4),

$$\mathbb{E}_0 \varphi^{\sigma_1} \le e^{\lambda \frac{n}{2} \log \frac{\sigma_0^2}{\sigma_1^2}} \Big(1 + \lambda\big(\frac{\sigma_0^2}{\sigma_1^2} - 1\big)\Big)^{-n/2}.$$

Now take $\lambda = 1/2$. Then using the fact that for every compact set $K \subset (0, \infty)$ there exists a constant $c > 0$ such that $(1/2)\log x - \log((1 + x)/2) \le -c(x - 1)^2$ for all $x \in K$, we find that there is a constant $\kappa_1 > 0$ such that the first inequality holds.

Next we note that for $Z = (Y - f)/\sigma$ and $h = (f - f_0)/\sigma_0$, Markov's inequality and (A.4) imply that

$$\begin{aligned}
&\mathbb{E}_{f,\sigma}(1 - \varphi^{\sigma_1}) \\
=&\mathbb{P}\Big(\frac{\sigma}{\sigma_0}\Big(\frac{\sigma_0^2}{\sigma_1^2} - 1\Big)\langle Z, h\rangle + \frac{1}{2}\frac{\sigma^2}{\sigma_0^2}\Big(\frac{\sigma_0^2}{\sigma_1^2} - 1\Big)\|Z\|^2 > \frac{n}{2}\log\frac{\sigma_0^2}{\sigma_1^2} - \frac{1}{2}\Big(\frac{\sigma_0^2}{\sigma_1^2} - 1\Big)\|h\|^2\Big)\Big) \\
&\le e^{-y} e^{\frac{\|v\|^2}{2(1+w)}} (1 + w)^{-n/2},
\end{aligned}$$

where

$$v_i = \frac{\sigma}{\sigma_0}\Big(\frac{\sigma_0^2}{\sigma_1^2} - 1\Big)h(x_i), \quad w = -\frac{\sigma^2}{\sigma_0^2}\Big(\frac{\sigma_0^2}{\sigma_1^2} - 1\Big),$$

$$y = \frac{n}{2}\log\frac{\sigma_0^2}{\sigma_1^2} - \frac{1}{2}\Big(\frac{\sigma_0^2}{\sigma_1^2} - 1\Big)\|h\|^2.$$

The terms without h in the exponent sum up to

$$-\frac{n}{2}\Big(\log\frac{\sigma_0^2}{\sigma_1^2} + \log\Big(1 - \frac{\sigma^2}{\sigma_0^2}\Big(\frac{\sigma_0^2}{\sigma_1^2} - 1\Big)\Big)\Big) = -\frac{n}{2}\Big(\log(1 + \delta_n) + \log\Big(1 - \frac{\sigma^2}{\sigma_0^2}\delta_n\Big)\Big),$$

where $\delta_n = \sigma_0^2/\sigma_1^2 - 1$. For $\delta_n \to 0$, Taylor's formula gives

$$\log(1 + \delta_n) + \log\Big(1 - \frac{\sigma^2}{\sigma_0^2}\delta_n\Big) = \Big(1 - \frac{\sigma^2}{\sigma_0^2}\Big)\delta_n - \frac{1}{2}\Big(1 + \frac{\sigma^4}{\sigma_0^4}\Big)\delta_n^2 + o(\delta_n^2).$$

If δ_n is small enough we have $1 + w \ge 1/2$ for large n. To complete the proof, also use that $\|v\|^2 \le n\delta_n^2(b^2/a^2)\varepsilon^2$. □

We can use the tests exhibited in the lemma to show that term II in (A.2) converges to 0 in $\mathbb{P}_0$-probability as $n \to \infty$. First we consider

$$\Pi_n\Big(\frac{\sigma^2}{\sigma_0^2} - 1 \le -M\varepsilon_n, \Big\|\frac{f-f_0}{\sigma_0}\Big\|_n \le M\varepsilon_n \,|\, Y\Big).$$

For any $\sigma_1 \in [a,b]$ the $\mathbb{E}_0$-expectation of this quantity is bounded by

$$\mathbb{E}_0\varphi_n^{\sigma_1} + \frac{1}{c_1}e^{(c_2+7))n\varepsilon_n^2} \sup_{\substack{\frac{\sigma^2}{\sigma_0^2}-1\le -M\varepsilon_n \\ \|(f-f_0)/\sigma_0\|_n \le M\varepsilon_n}} \mathbb{E}_{f,\sigma}(1-\varphi_n^{\sigma_1}) + \mathbb{P}_0(A_n^c)$$

Now take σ_1 such that $\sigma_0^2/\sigma_1^2 - 1 = \varepsilon_n$. Then by the lemma, the first term converges to 0 and the supremum in the second one is bounded by $\exp(-n\varepsilon_n^2(M/2 - \kappa_2 - \varepsilon_n M^2))$. Hence, the expression in the display converges to 0 for all large enough $M > 0$. The posterior probability

$$\Pi_n\Big(\frac{\sigma^2}{\sigma_0^2} - 1 \ge M\varepsilon_n, \Big\|\frac{f-f_0}{\sigma_0}\Big\|_n \le M\varepsilon_n \,|\, Y\Big)$$

can be handled similarly, by taking σ_1 such that $\sigma_0^2/\sigma_1^2 - 1 = -\varepsilon_n$.

A.2.4. Term III

We have

$$\mathbb{E}_0\Pi_n(f \in \mathcal{F}_n^c \,|\, Y) \le \mathbb{E}_0\Pi_n(f \in \mathcal{F}_n^c \,|\, Y)1_{A_n} + \mathbb{P}_0(A_n^c)$$

The second term converges to 0. By (A.3), Lemma A.2 and the prior mass assumption, the first term is bounded by $e^{7n\tilde\varepsilon_n^2}\Pi_n(f \in \mathcal{F}_n^c)$. By the second assumption of the theorem this converges to 0 as well.

References

Bhattacharya, A., Pati, D. and Dunson, D. B. (2012). Adaptive dimension reduction with a Gaussian process prior. *Preprint.*

Bickel, P. and Kleijn, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.* **40** 206–237.

Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.* **35** 2219–2232. MR2363969 (2009a:62179)

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287 (2010d:62069)

Castillo, I. (2012a). A semiparametric Bernstein - von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99.

Castillo, I. (2012b). Semiparametric Bernstein-von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya A* to appear.

De Blasi, P. and Hjort, N. L. (2009). The Bernstein-von Mises theorem in semiparametric competing risks models. *J. Statist. Plann. Inference* **139** 2316–2328. MR2507993 (2010h:62087)

De Jonge, R. and Van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.* **38** 3300–3320. MR2766853 (2012b:62136)

De Jonge, R. and Van Zanten, J. H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.* **6** 1984–2001.

Edmunds, D. E. and Triebel, H. (1996). *Function spaces, entropy numbers, differential operators*. *Cambridge Tracts in Mathematics* **120**. Cambridge University Press, Cambridge. MR1410258 (97h:46045)

Ghosal, S. and Van der Vaart, A. W. (2007). Convergence rates for posterior distributions for noniid observations. *Ann. Statist.* **35** 697–723.

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.* **32** 1556–1593. MR2089134 (2006m:62050)

Li, W. V. and Linde, W. (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability* **27** 1556–1578.

Lifshits, M. and Simon, T. (2005). Small deviations for fractional stable processes. *Ann. Inst. H. Poincaré Probab. Statist.* **41** 725–752. MR2144231 (2006d:60081)

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

Rivoirard, V. and Rousseau, J. (2012). Bernstein–Von Mises Theorem for linear functionals of the density. *Ann. Statist.* to appear.

Schumaker, L. L. (1981). *Spline functions: basic theory*. John Wiley & Sons Inc., New York. Pure and Applied Mathematics, A Wiley-Interscience Publication. MR606200 (82j:41001)

Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.* **97** 222–235. MR1947282 (2003i:62029)

Tokdar, S. A. (2011). Dimension adaptability of Gaussian process models with variable selection and projection. *Preprint*.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge University Press, Cambridge. MR1652247 (2000c:62003)

Van der Vaart, A. W. and Van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663 (2009i:62068)

Van der Vaart, A. W. and Van Zanten, J. H. (2008b). *Reproducing Kernel Hilbert Spaces of Gaussian priors*. *IMS Collections* **3** 200–222. Institute of Mathematical Statistics.

Van der Vaart, A. W. and Van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442 (2010j:62105)

Van der Vaart, A. W. and Van Zanten, J. H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12** 2095–2119. MR2819028

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. *Springer Series in Statistics*. Springer-Verlag, New York. With applications to statistics. MR1385671 (97g:60035)

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372. MR522220 (80f:62047)