Journal of Econometrics 185 (2015) 409–419

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian regression with nonparametric heteroskedasticity

Andriy Norets

Department of Economics, Brown University, United States

ARTICLE INFO

Article history:
Received 3 March 2014
Received in revised form
1 September 2014
Accepted 19 December 2014
Available online 7 January 2015

Keywords:
Bayesian linear regression
Heteroskedasticity
Misspecification
Posterior consistency
Semiparametric Bernstein–von Mises theorem
Semiparametric efficiency
Gaussian process priors
Multivariate Bernstein polynomials

ABSTRACT

This paper studies large sample properties of a semiparametric Bayesian approach to inference in a linear regression model. The approach is to model the distribution of the regression error term by a normal distribution with the variance that is a flexible function of covariates. The main result of the paper is a semiparametric Bernstein–von Mises theorem under misspecification: even when the distribution of the regression error term is not normal, the posterior distribution of the properly recentered and rescaled regression coefficients converges to a normal distribution with the zero mean and the variance equal to the semiparametric efficiency bound.

1. Introduction

A linear model $Y_i = X_i'\beta_0 + \epsilon_i$ with the conditional moment restriction $E(\epsilon_i|X_i) = 0$ is a standard regression model, which is widely used in statistics and econometrics. This paper analyzes asymptotic properties of a Bayesian semiparametric approach to estimation of this model. The approach is to model the distribution of the error term by a normal distribution with the variance that is a flexible function of covariates. For example, Gaussian process priors, splines, or polynomials can be used to build a prior for the variance. Normality of the error term guarantees that the Kullback–Leibler (KL) distance between the model and the data generating process (DGP), which does not necessarily satisfy the normality assumption, is minimized at the data generating values of the linear coefficients and the conditional variance of the error term. Thus, one can expect that the posterior asymptotically concentrates around the true values for these two parameters. The normality assumption can also be justified by appealing to the principle of maximum entropy of Jaynes (1957) when only the first two conditional moments are of interest.

The main result of the paper is a semiparametric Bernstein–von Mises theorem under misspecification: even when the distribution of the regression error term is not normal in the DGP, the posterior distribution of the properly recentered and rescaled regression coefficients converges to a normal distribution with the zero mean and the variance equal to the semiparametric efficiency bound. The equality of the variance to the semiparametric efficiency bound suggests that the Bayesian inference about the linear coefficients based on this model is conservative in the following sense: the posterior variance in a correctly specified parametric model is likely to be smaller than the posterior variance in a model that postulates normally distributed errors with the flexibly modeled variance. With carefully specified priors, Bayesian procedures usually behave well in small samples. Thus, the Bayesian normal linear regression with nonparametric heteroskedasticity can also be an attractive alternative to classical semiparametrically efficient estimators from Carroll (1982) and Robinson (1987). At the same time, the results of the paper provide a Bayesian interpretation to these classical estimators.

Several different approaches to inference in a regression model have been proposed in the Bayesian framework. In a standard textbook linear regression model, normality of the error terms is assumed. More recent literature relaxed the normality assumption by using mixtures of normal or Student t distributions. However, if the shape of the error distribution depends on covariates then the posterior may not concentrate around the data generating values of the linear coefficients (Müller, 2013). Lancaster (2003) and Poirier (2011) do not assume linearity of the regression function and treat the linear projection coefficients as the parameters of interest. They use Bayesian bootstrap (Rubin, 1981) to justify from the Bayesian perspective the use of the ordinary least squares es-

E-mail address: andriy_norets@brown.edu.

http://dx.doi.org/10.1016/j.jeconom.2014.12.006
0304-4076/

timator with a heteroskedasticity robust covariance matrix. Pelenis (2014) demonstrates posterior consistency in a semiparametric model with a parametric specification for the regression function and a nonparametric specification for the conditional distribution of the regression error term. It is also possible to estimate a fully nonparametric model for the distribution of the response conditional on covariates, see, for example, Peng et al. (1996), Wood et al. (2002), Geweke and Keane (2007), Villani et al. (2009), and Norets (2010) for Bayesian models based on smoothly mixing regressions or mixtures of experts and MacEachern (1999), DeIorio et al. (2004), Griffin and Steel (2006), Dunson and Park (2008), Chung and Dunson (2009), Norets and Pelenis (2014), and Pati et al. (2013) for models based on dependent Dirichlet processes. These fully nonparametric models require a lot of data for reliable estimation results and prior specification is nontrivial. The model considered in the present paper is more parsimonious. Nevertheless, it delivers consistent estimation of the first two conditional moments, conservative inference about the regression coefficients, and it is robust to misspecification of the regression error distribution. Thus, it can be thought of as a useful intermediate step between fully nonparametric and oversimplistic models.

Bayesian Markov chain Monte Carlo (MCMC) estimation algorithms for the normal regression with flexibly modeled variance have been developed in the literature; see, for example, Yau and Kohn (2003) and Chib and Greenberg (2013), who use transformed splines, or Goldberg et al. (1998), who use transformed Gaussian process prior for modeling the variance. In those papers, the models with flexibly modeled variances were shown to perform well in simulation studies. Thus, the present paper considers only the theoretical properties of the model.

The rest of the paper is organized as follows. Section 2 describes the DGP. The model is described in Section 3. The Bernstein–von Mises theorem is presented in Section 4. The assumptions of the theorem are verified in Section 5 for priors based on truncated Gaussian processes and multivariate Bernstein polynomials. Section 6 concludes. Proofs are delegated to Section 7.

2. Data generating process

The data are assumed to include n observations on a response variable and covariates $(\mathbf{Y}^n, \mathbf{X}^n) = (Y_1, \ldots, Y_n, X_1, \ldots, X_n)$, where $Y_i \in \mathcal{Y} \subset R$ and $X_i \in \mathcal{X} \subset \mathbb{R}^d$, $i \in \{1, \ldots, n\}$. $\mathcal{X}$ is assumed to be convex and bounded set with a nonempty interior. The observations are independently identically distributed (iid), $(Y_i, X_i) \sim F_0$. The joint DGP distribution F_0 is assumed to have a conditional density $f_0(Y_i|X_i)$ with respect to (w.r.t.) the Lebesgue measure. The distribution of the infinite sequence of observations, $(\mathbf{Y}^\infty, \mathbf{X}^\infty)$, is denoted by F_0^∞. Hereafter, the expectations $E(\cdot)$ and $E(\cdot|\cdot)$ are taken w.r.t. the DGP F_0^∞.

Let us make the following assumptions about the data generating process. First, $E(Y_i|X_i) = X_i'\beta_0$. Thus, for $\epsilon_i = Y_i - X_i'\beta_0$, $E(\epsilon_i|X_i) = 0$. Second, $E(X_iX_i')$ is invertible. Third, $\sigma_0^2(x) = E(\epsilon_i^2|X_i = x)$ is well defined for any $x \in \mathcal{X}$, and for some $0 < \underline{\sigma} < \overline{\sigma} < \infty$,

$$\sigma_0(x) \in (\underline{\sigma}, \overline{\sigma}), \quad \forall\, x \in \mathcal{X}. \tag{1}$$

3. Model, prior, and pseudo true parameter values

In the model, it is assumed that the regression error term is normally distributed conditional on covariates, $\epsilon_i|X_i, \beta, \sigma \sim N(0, \sigma^2(X_i))$. Note that this normality assumption is not made for the DGP and the model can be misspecified.

Let $S \subset \{\sigma : \mathcal{X} \to [\underline{\sigma}, \overline{\sigma}]\}$ be a complete separable metric space and $\mathcal{A}$ be a Borel σ-field on $\mathbb{R}^d \times S$. Let Π denote a prior distribution for (β, σ) on $(\mathbb{R}^d \times \mathcal{S}, \mathcal{A})$. Π is a product of a normal $N(\underline{\beta}, \underline{H}^{-1})$ prior for β truncated to $[-B, B]^d$, where a lower bound on a finite constant B is specified below, and a prior distribution for σ on S. It is also assumed that $\sigma_0 \in S$.

The prior on S will be assumed to put a sufficiently large probability on the following class of smooth functions

$$S_{M,\alpha} = \Big\{\sigma : \mathcal{X} \to [\underline{\sigma}, \overline{\sigma}] : \max_{k_1+\cdots+k_d \le \underline{\alpha}} \sup_{x \in \mathcal{X}} |\partial^k \sigma(x)| + \max_{k_1+\cdots+k_d = \underline{\alpha}} \sup_{x \ne z \in \mathcal{X}} \frac{|\partial^k \sigma(x) - \partial^k \sigma(z)|}{\|x - z\|^{\alpha - \underline{\alpha}}} \le M \Big\},$$

where $k = (k_1, \ldots, k_d)$ is a multi-index, $\partial^k = \partial^k / \partial^{k_1} x_1 \cdots \partial^{k_d} x_d$ is a partial derivative operator, and $\underline{\alpha}$ is the greatest integer strictly smaller than $\alpha > 0$.

The distribution of covariates is assumed to be ancillary and it is not modeled. The likelihood function is given by

$$p(\mathbf{Y}^n|\mathbf{X}^n, \beta, \sigma) = \prod_{i=1}^n p_{\beta,\sigma}(Y_i|X_i),$$

$$p_{\beta,\sigma}(Y_i|X_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma(X_i)} \exp\left(-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2(X_i)}\right).$$

For $A \in \mathcal{A}$, the posterior is given by

$$\Pi(A|\mathbf{Y}^n, \mathbf{X}^n) = \frac{\int_A p(\mathbf{Y}^n|\mathbf{X}^n, \beta, \sigma) d\Pi(\beta, \sigma)}{\int_{\mathbb{R}^d \times S} p(\mathbf{Y}^n|\mathbf{X}^n, \beta, \sigma) d\Pi(\beta, \sigma)}.$$

In misspecified models, parameter values minimizing the KL distance between the model and the DGP are called pseudo true parameter values. It is well known that in models with finite dimensional parameters the maximum likelihood and Bayesian estimators are consistent for the pseudo true parameter values under weak regularity conditions (see Huber (1967), White (1982), and Gourieroux et al. (1984) for classical results and Geweke (2005) and Kleijn and van der Vaart (2012) for Bayesian results). Analogous results for misspecified infinite dimensional models are obtained in Kleijn and van der Vaart (2006).[1] Thus, the following lemma suggests that in the regression model described above the posterior concentrates around (β_0, σ_0) in large samples.

Lemma 1. *Consider the DGP and the model described above. Suppose* $E(|\log f_0(Y_i|X_i)|) < \infty$. *Then,*

$$(\beta_0, \sigma_0) \in \operatorname*{argmin}_{\beta \in R^d, \sigma : \mathcal{X} \to [\underline{\sigma}, \overline{\sigma}]} E\left(\log \frac{f_0(Y_i|X_i)}{p_{\beta,\sigma}(Y_i|X_i)}\right).$$

If $E(X_iX_i')$ *is positive definite, then the minimizer is* F_0 *almost surely unique.*

The lemma is proved in Section 7.

4. Semiparametric Bernstein–von Mises theorem

The standard Bernstein–von Mises theorem shows that in well behaved parametric models the posterior distribution centered at an efficient estimator and scaled by $\sqrt{n}$ converges to a normal distribution with the zero mean and the variance equal to the inverse of the Fisher information, see van der Vaart (1998) for a textbook treatment under weak regularity conditions. Thus,

[1] There is a considerable body of research on posterior consistency in correctly specified nonparametric models. A general weak posterior consistency theorem for density estimation was established by Schwartz (1965). Barron (1988), Barron et al. (1999), and Ghosal et al. (1999) developed theory of strong posterior consistency. An alternative approach to consistency was introduced by Walker (2004). Posterior convergence rates were studied in Ghosal et al. (2000) and Shen and Wasserman (2001). Belitser and Ghosal (2003), Ghosal et al. (2003), Huang (2004), Scricciolo (2006), Ghosal et al. (2008), van der Vaart and van Zanten (2009), Kruijer et al. (2010), and Gine and Nickl (2011) among others analyzed adaptation of posterior convergence rates.

the theorem implies asymptotic equivalence between standard confidence and credible sets. Chernozhukov and Hong (2003) and Kleijn and van der Vaart (2012) prove Bernstein–von Mises type theorems for misspecified parametric models,[2] and Panov and Spokoiny (2013) consider settings with increasing dimension of the nuisance parameter. Shen (2002) gave a set of conditions for asymptotic normality of the posterior of a finite dimensional part of the parameter in semiparametric models. The conditions are general but difficult to verify. Deriving easier to verify sufficient conditions for the non- and semiparametric Bernstein–von Mises theorem is an active area of current research, see, for example, Bickel and Kleijn (2012), Castillo (2012), Rivoirard and Rousseau (2012), Kleijn and Knapik (2012), Kato (2013), Castillo and Nickl (2013), Castillo and Rousseau (2013), and Castillo and Nickl (2014). Misspecified semiparametric models are not covered by the existing results.

Frequentist estimation of a semiparametric regression outlined in Section 2 was considered in Chamberlain (1987). He shows that the semiparametric efficiency bound for estimation of β_0 is given by $\left(E(X_iX_i'\sigma_0(X_i)^{-2})\right)^{-1}$. This is the asymptotic variance of the generalized least squares estimator under known σ_0,

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{n} \frac{X_iX_i'}{\sigma_0(X_i)^2}\right)^{-1} \sum_{i=1}^{n} \frac{X_iY_i}{\sigma_0(X_i)^2}.$$

It follows from Carroll (1982) and Robinson (1987) that if σ_0 is estimated by kernel smoothing or nearest neighbor methods and plugged in the formula for $\hat{\beta}_{GLS}$ the resulting estimator attains the efficiency bound. The following Bernstein–von Mises theorem is a Bayesian analog of these results.

Theorem 1. *For the DGP from Section 2 and the prior from Section 3, let us make the following additional assumptions.*

1. $E[\epsilon_i^4] < \infty$.
2. *There exist* $\alpha > d/2$ *and positive sequences* δ_n *and* M_n *satisfying*
$$\delta_n^4 n \to \infty \quad \text{and} \quad (M_n/\delta_n^2)^{d/\alpha}/(\delta_n^4 n) \to 0,$$
$$\sqrt{n}E\left(|X_{ij}\epsilon_i| 1\{|X_{ij}\epsilon_i| > C\sqrt{n}\delta_n^{(1+d/(2\alpha))} M_n^{-d/(2\alpha)}\}\right) \to 0,$$
$$\forall\, C \in (0,\infty),\ j = 1,\dots,d,$$
$$\delta_n^{(1-d/(2\alpha))} M_n^{d/(2\alpha)} \to 0.$$
3. *For any constant* $A > 0$, $\Pi(\sigma : \sigma \in S^c_{M_n,\alpha}) \exp\{An\} \to 0$ *as* $n \to \infty$.
4. *The truncation bound in the prior for* β *satisfies* $B > (\overline{\sigma}^2/\underline{\sigma}^2)\, \|E|X_iY_i|\|_2/e_{\min}(E(X_iX_i'))$, *where* $e_{\min}(E(X_iX_i'))$ *is the smallest eigenvalue of* $E(X_iX_i')$.
5. *For any* $C > 0$ *there exists* $r \in (0,1)$ *such that for all sufficiently large* n,
$$\Pi\left(\sigma, \beta :\ E\left[\log \frac{p_{\beta_0,\sigma_0}(Y_i|X_i)}{p_{\beta,\sigma}(Y_i|X_i)}\right] \le C\delta_n^2\right) \ge \exp\{-nrC\delta_n^2\}.$$

Then, the total variation distance

$$d_{TV}\left(\Pi[\sqrt{n}(\beta - \hat{\beta}_{GLS})|\mathbf{Y}^n, \mathbf{X}^n], N\left(0, \left(E[X_iX_i'\sigma_0(X_i)^{-2}]\right)^{-1}\right)\right) \quad (2)$$

converge to 0 *in* F_0^∞ *probability. This result also holds if* $\hat{\beta}_{GLS}$ *in* (2) *is replaced by the posterior mean* $\int \beta d\Pi(\beta|\mathbf{Y}^n, \mathbf{X}^n)$.

The theorem is proved in Section 7. The proof exploits the fact that the conditional posterior for $\sqrt{n}(\beta - \hat{\beta}_{GLS})$ given σ is a truncated normal distribution, which is close to $N\left(0, \left(E[X_iX_i'\sigma_0(X_i)^{-2}]\right)^{-1}\right)$ when σ is close to σ_0. Arguments based on maximal inequalities for classes of functions that change with n from Pollard (1984) and van der Vaart (1998) are used to make this claim precise and to show that the posterior of σ concentrates on σ_0 sufficiently fast.

The assumption of the bounded parameter space ($\underline{\sigma} < \sigma_0 < \overline{\sigma}$ and $B < \infty$) is admittedly restrictive. However, it appears to be difficult to relax under misspecification. Similar assumptions are also made in the previous related literature: Kleijn and van der Vaart (2006) assumed a fixed variance and a bounded range for the conditional mean in their misspecified regression example; Carroll (1982) and Andrews (1994) assume a strictly positive lower bound for the conditional variance function. Bickel and Kleijn (2012) do not assume bounded space for regression coefficients in their analysis of homoskedastic partially linear regression, which suggest that it could be possible to use their approach to relax this assumption under a correctly specified model.

The assumptions of Theorem 1 on the prior are easy to verify in applications. Assumption 5 of prior concentration on KL neighborhoods is standard in the literature. Verification of Assumption 2 can be simplified when the distribution of ϵ_i has subexponential tails: when $f_0(x'\beta_0 \pm \epsilon|x) \le e^{-D\epsilon}$ for some $D > 0$, all $x \in \mathcal{X}$, and all sufficiently large ϵ, the assumption holds for

$$M_n = n^{\gamma_1}, \qquad \delta_n = n^{-\gamma_2},$$
$$\gamma_2 < \frac{1}{6 + d/\alpha},\ \gamma_1 < [2\alpha/d - 1]\gamma_2. \quad (3)$$

Assumption 3 is easy to verify for priors that provide explicit distributions for derivatives. The following section verifies the assumptions of the theorem for priors based on truncated Gaussian processes and multivariate Bernstein polynomials.

5. Examples of priors

5.1. Truncated Gaussian process

Priors based on Gaussian processes are extensively used in Bayesian nonparametrics; see, for example, Tokdar and Ghosh (2007), Tokdar (2007), van der Vaart and van Zanten (2008), Liang et al. (2009), and Tokdar et al. (2010). In this section, I consider a prior for σ based on integrated Brownian motion. Relevant technical background can be found in Section 4 of van der Vaart and van Zanten (2008).

Let $\mathcal{X} = [0,1]$, $W(x)$ be a Brownian motion on $\mathcal{X}$ with continuous sample paths, $(I^1W)(x) = \int_0^x W(t)dt$, and $(I^jW)(x) = \int_0^x (I^{j-1}W)(t)dt$ for $j \ge 2$. Suppose $\sigma_0 \in S_{M,\alpha_0}$ for some (M,α_0) and $\underline{\sigma} < \sigma_0 < \overline{\sigma}$. For a positive integer J, let us model $\sigma(x)$ by $(I^JW)(x) + \sum_{j=0}^J Z_j x^j/j!$ truncated to $[\underline{\sigma}, \overline{\sigma}]$, where Z_j's are i.i.d. $\mathcal{N}(0,1)$ independent of W. More formally, let $\mathcal{W}$ denote the probability measure for W and Z_j's, which is defined on $C([0,1])$ and its Borel σ-field. Define an event,

$$T = \left\{(I^JW)(x) + \sum_{j=0}^{J} Z_j x^j/j! \in [\underline{\sigma}, \overline{\sigma}],\ \forall x \in [0,1]\right\}$$

and for a measurable set E,

$$\Pi(\sigma \in E) = \mathcal{W}\left(\left[(I^kW)(x) + \sum_{j=0}^{J} Z_j x^j/j! \in E\right] \bigcap T\right) \Big/ \mathcal{W}(T).$$

Proposition 1. *Assume that the tails of* ϵ_i *are subexponential. Then, for any* $J \ge 3$ *and* $\alpha_0 \in ((J+1)/(2J-1), J+1/2]$, *Assumptions* 2, 3 *and* 5 *of Theorem 1 hold.*

5.2. Multivariate Bernstein polynomials

Priors based on Bernstein polynomials and related mixtures of beta models were considered in Petrone (1999), Ghosal (2001),

[2] In parametric misspecified models, asymptotic normality of posterior does not in general imply asymptotic equivalence of standard confidence and credible sets due to the failure of the information equality.

Petrone and Wasserman (2002), Kruijer and van der Vaart (2008), Rousseau (2010), and Burda and Prokhorov (2013) among others. Bernstein polynomials play a prominent role in function approximation literature especially when it comes to shape preserving approximation, see a monograph by Lorentz (1986). As shown below, multivariate Bernstein polynomials are also convenient for setting up priors that respect bounds on partial derivatives and thus satisfy sufficient conditions in Theorem 1.

In this subsection, $\mathcal{X} = [0,1]^d$. For functions $f : [0,1]^d \to \mathbb{R}$, a Bernstein polynomial operator of order m_i w.r.t. coordinate i is defined as follows,

$$(B_i^{m_i} f)(x) = \sum_{j_i=0}^{m_i} f\left(\frac{j_i}{m_i}, x_{-i}\right) \binom{m_i}{j_i} x_i^{j_i} (1-x_i)^{m_i - j_i},$$

where $x = (x_i, x_{-i})$, $x_i \in [0,1]$, and $x_{-i} \in [0,1]^{d-1}$. Also let $\partial_i^{\lambda_i}$ stand for a partial derivative of order λ_i w.r.t. to coordinate i and $D_i^{m_i}$ stand for a first difference operator[3] w.r.t. coordinate i

$$D_i^{m_i} f(x) = \frac{f((1-1/m_i)x_i + 1/m_i, x_{-i}) - f((1-1/m_i)x_i, x_{-i})}{1/m_i}.$$

For a given m_i, define λ_ith difference by $D_i^{m_i,\lambda_i} = D_i^{m_i-\lambda_i+1} D_i^{m_i-\lambda_i+2} \cdots D_i^{m_i}$. For multi-indices $m = (m_1, \ldots, m_k)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$, $m \geq \lambda$, denote a multivariate Bernstein polynomial operator of order m by $B^m = B_1^{m_1} \cdots B_k^{m_k}$, a multivariate difference operator by $D^{m,\lambda} = D_1^{m_1,\lambda_1} \cdots D_k^{m_k,\lambda_k}$, and a partial derivative of order λ by $\partial^\lambda = \partial_1^{\lambda_1} \cdots \partial_k^{\lambda_k}$.

Part (v) of Lemma 11 shows that

$$\partial_i^1 B_i^{m_i} = B_i^{m_i - 1} D_i^{m_i}. \tag{4}$$

From a repeated application of (4) it follows that $\partial_i^{\lambda_i} B_i^{m_i} = B_i^{m_i-\lambda_i} D_i^{m_i,\lambda_i}$. Since for any m_i, q, and $i \neq k$, we have $B_i^{m_i} D_k^q = D_k^q B_i^{m_i}$, it follows that

$$\partial^\lambda B^m = B^{m-\lambda} D^{m,\lambda}. \tag{5}$$

For $m \in \mathbb{N}^d$, let us define a collection of points (a grid on $[0,1]^d$)

$$G^m = \{x_j \in [0,1]^d : x_j = j/m = (j_1/m_1, \ldots, j_k/m_k),\ o \leq j \leq m\},$$

where the inequalities hold coordinate-wise and o is a vector of zeros. Let us denote a multivariate Bernstein polynomial of order m with parameter $g = \{g_j, o \leq j \leq m\}$ by

$$B^m(x; g) = \sum_{o \leq j \leq m} g_j \prod_{i=1}^d \binom{m_i}{j_i} x_i^{j_i} (1-x_i)^{m_i - j_i}.$$

For any given parameter g and a function $f : [0,1]^d \to \mathbb{R}$ such that $g = f(G^m)$,

$$B^m(x; g) = (B^m f)(x), \quad \forall\, x \in [0,1]^d,$$

and we can also define

$$D^{m,\lambda}(x; g) = (D^{m,\lambda} f)(x), \quad \forall\, x \in G^{m-\lambda}. \tag{6}$$

Let us consider a sample size dependent prior[4] Π_n that puts probability 1 on m such that $m_1 = \cdots = m_d = n^{\gamma_3}$, where

$$\gamma_3 = \frac{4 + d/\alpha}{d[6 + d/\alpha]}, \quad \alpha = \lfloor d/2 \rfloor + 1, \tag{7}$$

and the dependence of m on n is not reflected in the notation for brevity.

Conditional on m, $\sigma(x) = B^m(x; g^m)$ and the prior for g^m is assumed to have a positive density w.r.t. the Lebesgue measure on $\mathbb{R}^{(m_1+1)^d}$. This density is assumed to be bounded from below by $\underline{\pi}_g^{(m_1+1)^d}$ for some $\underline{\pi}_g > 0$ on the restrictions defined in (8) and equal to zero elsewhere.

$$\begin{aligned} &g_j^m \in [\underline{\sigma}, \overline{\sigma}] \quad \text{for } o \leq j \leq m; \\ &|D^{m,\lambda}(x; g^m)| \leq M_n \quad \text{for } \sum_{i=1}^d \lambda_i \leq \lfloor d/2 \rfloor + 1, x \in G^{m-\lambda}, \end{aligned} \tag{8}$$

$$M_n = n^{\gamma_1}, \quad \gamma_1 = \frac{[2\alpha/d - 1](1-s)^2}{d[6 + d/\alpha]}, \tag{9}$$

and s is a constant in $(0,1)$.

Lemma 11 provides explicit formulas for $D^{m,\lambda}(G^{m-\lambda}; g^m)$ that are linear in g^m. Thus, the model with the sample size dependent prior can be implemented by truncating a positive density for the Bernstein polynomial coefficients to a set of linear inequality restrictions in (8) and using Markov chain Monte Carlo methods for estimation.

Proposition 2. *Assume that the tails of ϵ_i are subexponential and σ_0 has uniformly bounded partial derivatives up to order $\lfloor d/2 \rfloor + 2$. Then, for the prior specified above, Assumptions 2, 3 and 5 of Theorem 1 hold.*

5.3. Other possible priors

It should be possible to verify the assumptions of Theorem 1 for other common nonparametric prior distributions. First, a wavelet prior in the spirit of Gine and Nickl (2011) (p. 5 display (5)) is likely to satisfy the assumptions as it imposes an explicit uniform bound on the Hölder norm. Such a prior might be easier to implement than the prior of Section 5.1 since it does not involve truncation (the advantage of the prior of Section 5.1 is that it does not require uniform bounds on derivatives). Second, the sufficient conditions for the truncated Gaussian process in Proposition 1 show that an increase in the smoothness of the process, J, only weakens the sufficient conditions. Thus, truncated Gaussian processes with analytical sample paths, such as a process with the squared-exponential covariance kernel, are likely to satisfy the assumptions as well. Finally, a verification of the assumptions for a prior based on splines was present in an earlier version of the paper and is omitted from the current version for brevity.

6. Conclusion

The paper proves asymptotic normality of the posterior of the coefficients in possibly misspecified heteroskedastic linear regression model. The model is shown to be robust to misspecification in the distribution of the regression error term. Thus, this model should be a more prominent part of the Bayesian toolbox for regression analysis.

7. Proofs

Proof. Lemma 1.

$$\int \log \frac{f_0(y|x)}{(2\pi)^{-0.5}\sigma(x)^{-1} \exp\{-0.5(y - x'\beta)^2/\sigma(x)^2\}} dF_0(y,x)$$

$$= \int \left[\log f_0(y|x) + 0.5 \log(2\pi\sigma(x)^2) + \frac{(y - x'\beta)^2}{2\sigma(x)^2}\right] dF_0(y,x).$$

Since $E[(Y_i - X_i'\beta)^2|X_i) = \sigma_0(X_i)^2 + [X_i'(\beta - \beta_0)]^2$, $\beta = \beta_0$ is a minimizer of the KL distance for any $\sigma : \mathcal{X} \to [\underline{\sigma}, \overline{\sigma}]$. It is a unique minimizer if $E(X_i X_i') > 0$. For any fixed x, $\log \sigma^2(x) + \sigma_0^2(x)/\sigma^2(x)$ is uniquely minimized at $\sigma(x) = \sigma_0(x)$. □

[3] This definition is from Majer (2012), who briefly outlines an argument for why partial derivatives of multivariate Bernstein polynomials for f approximate corresponding partial derivatives of f; the details of the argument are worked out here in Lemma 11.

[4] The proof of Theorem 1 does not require any changes when the prior is indexed by n.

Proof. Theorem 1.

Conditional on σ, the posterior of β, $\Pi(\beta|\sigma, \mathbf{Y}^n, \mathbf{X}^n)$, is $N(\overline{\beta}, H^{-1})$ truncated to $[-B, B]^d$, where

$$H = \underline{H} + \sum_i \frac{X_i X_i'}{\sigma(X_i)^2} \quad \text{and} \quad \overline{\beta} = H^{-1}\left(\underline{H}\underline{\beta} + \sum_i \frac{X_i Y_i}{\sigma(X_i)^2}\right).$$

A derivation of the conditional posteriors in linear regression models can be found, for example, in Geweke (2005). The (marginal) posterior of β can be expressed as

$$\Pi(\beta|\mathbf{Y}^n, \mathbf{X}^n) = \int \Pi(\beta|\sigma, \mathbf{Y}^n, \mathbf{X}^n) d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n).$$

The conditional posterior of the normalized coefficients $z = \sqrt{n}(\beta - \hat{\beta}_{GLS})$ is

$$\Pi(z|\sigma, \mathbf{Y}^n, \mathbf{X}^n) \propto \phi\left(z, \sqrt{n}(\overline{\beta} - \hat{\beta}_{GLS}), (H/n)^{-1}\right) \times 1_{\sqrt{n}([-B,B]^d - \hat{\beta}_{GLS})}(z),$$

where $\phi(\cdot, \cdot, \cdot)$ denotes the density of the normal distribution and $1_{\cdot}(\cdot)$ is an indicator function.

The total variation distance of interest can be expressed as follows

$$\begin{aligned} & d_{TV}\left(\Pi[z|\mathbf{Y}^n, \mathbf{X}^n], N\left(0, \left(E[X_i X_i' \sigma_0(X_i)^{-2}]\right)^{-1}\right)\right) \\ &= \int \left| \int \Pi(z|\sigma, \mathbf{Y}^n, \mathbf{X}^n) d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n) \right. \\ &\quad \left. - \phi\left(z, 0, \left(E[X_i X_i' \sigma_0(X_i)^{-2}]\right)^{-1}\right) \right| dz \\ &\leq \int \int \left| \Pi(z|\sigma, \mathbf{Y}^n, \mathbf{X}^n) \right. \\ &\quad \left. - \phi\left(z, 0, \left(E[X_i X_i' \sigma_0(X_i)^{-2}]\right)^{-1}\right) \right| dz d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n). \end{aligned} \tag{10}$$

At this point in the proof, let us derive a bound on the total variation distance between two normal distributions and introduce notation for various matrix norms that will be used below. By Kemperman (1969) (or Proposition 1.2.2 in Ghosh and Ramamoorthi (2003)), the total variation distance is bounded by 2 times the square root of the KL distance. The KL distance between two normal distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ is equal to

$$\begin{aligned} & \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1 - I) + (\mu_1 - \mu_2)'\Sigma_2^{-1}(\mu_1 - \mu_2)\right) \\ &\leq \frac{\left||\Sigma_2^{-1}| - |\Sigma_1^{-1}|\right|}{\min(|\Sigma_2^{-1}|, |\Sigma_1^{-1}|)} + d \cdot \|\Sigma_2^{-1} - \Sigma_1^{-1}\|_\infty \cdot \|\Sigma_1\|_\infty \\ &\quad + \|\mu_1 - \mu_2\|_2^2 \cdot \|\Sigma_2^{-1}\|_2, \end{aligned} \tag{11}$$

where $|\Sigma|$ denotes the determinant of Σ, a matrix norm $\|\Sigma\|_\infty = \max_{ij} |[\Sigma]_{ij}|$ is the largest element of Σ in the absolute value, and $\|\Sigma\|_2 = \sup_\mu \|\Sigma\mu\|_2/\|\mu\|_2$ is a matrix norm induced by the standard norm on $\mathbb{R}^d$, $\|\mu\|_2^2 = \sum_{i=1}^d \mu_i^2$.

By Lemma 2 and inequality (11),

$$\begin{aligned} & d_{TV}\left(\Pi[\sqrt{n}(\beta - \hat{\beta}_{GLS})|\mathbf{Y}^n, \mathbf{X}^n], N\left(0, \left(E[X_i X_i' \sigma_0(X_i)^{-2}]\right)^{-1}\right)\right) \\ &\leq \int \left[2\sqrt{A_n + B_n + C_n}/D_n + (1 - D_n)/D_n + 1 - E_n\right] \\ &\quad \times d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n), \end{aligned}$$

where

$$A_n = \frac{\left||H/n| - |E[X_i X_i' \sigma_0(X_i)^{-2}]|\right|}{\min(|H/n|, |E[X_i X_i' \sigma_0(X_i)^{-2}]|)},$$

$$\begin{aligned} B_n &= d \cdot \|H/n - E[X_i X_i' \sigma_0(X_i)^{-2}]\|_\infty \cdot \|\left(E[X_i X_i' \sigma_0(X_i)^{-2}]\right)^{-1}\|_\infty, \\ C_n &= \|H/n\|_2 \cdot \left\| \left(\frac{1}{n}\sum_i \frac{X_i X_i'}{\sigma_0(X_i)^2}\right)^{-1} \frac{1}{\sqrt{n}} \sum_i \frac{X_i Y_i}{\sigma_0(X_i)^2}, \right. \\ &\quad \left. - (H/n)^{-1}\left(\frac{\underline{H}\underline{\beta}}{\sqrt{n}} + \frac{1}{\sqrt{n}}\sum_i \frac{X_i Y_i}{\sigma(X_i)^2}\right)\right\|_2^2, \\ D_n &= \int_{\sqrt{n}([-B,B]^d - \hat{\beta}_{GLS})} \phi\left(z, \sqrt{n}(\overline{\beta} - \hat{\beta}_{GLS}), (H/n)^{-1}\right) dz, \\ E_n &= \int_{\sqrt{n}([-B,B]^d - \hat{\beta}_{GLS})} \phi\left(z, 0, \left(E[X_i X_i' \sigma_0(X_i)^{-2}]\right)^{-1}\right) dz. \end{aligned}$$

It is shown in Lemma 3 that E_n and D_n have lower bounds that do not depend on σ and converge to 1 in F_0^∞ probability. By Lemmas 4–6, which bound (A_n, B_n, C_n), and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any nonnegative a and b, to prove the theorem it suffices to show that $\int \|\sum_i X_i \epsilon_i (\sigma_0(X_i)^{-2} - \sigma(X_i)^{-2})/\sqrt{n}\|_2 d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n)$ and $\int d_2(\sigma_0^{-2}, \sigma^{-2}) d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n)$ converge to zero in F_0^∞ probability, where $d_2(\sigma^{-2}, \sigma_0^{-2}) = (\int [\sigma^{-2}(x) - \sigma_0^{-2}(x)]^2 dF_0(x))^{0.5}$.

$$\begin{aligned} & \int \left\| \sum_i X_i \epsilon_i (\sigma_0(X_i)^{-2} - \sigma(X_i)^{-2})/\sqrt{n} \right\|_2 d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n) \\ &\leq \sup_{\sigma \in S} \left\| \sum_i X_i \epsilon_i (\sigma_0^{-2}(X_i) - \sigma^{-2}(X_i))/\sqrt{n} \right\|_2 \\ &\quad \cdot \left[\Pi(d_2(\sigma_0^{-2}, \sigma^{-2}) > \delta_n \cap S_{M_n,\alpha}|\mathbf{Y}^n, \mathbf{X}^n) \right. \\ &\quad \left. + \Pi(S_{M_n,\alpha}^c|\mathbf{Y}^n, \mathbf{X}^n)\right] + \sup_{\{\sigma \in S_{M_n,\alpha} : d_2(\sigma_0^{-2}, \sigma^{-2}) \leq \delta_n\}} \left\| \sum_i X_i \epsilon_i \right. \\ &\quad \left. \times (\sigma_0^{-2}(X_i) - \sigma^{-2}(X_i))/\sqrt{n} \right\|_2. \end{aligned} \tag{12}$$

The right hand side of the above display converges to zero in F_0^∞ outer probability[5] if

$$\sqrt{n} \cdot \Pi(d_2(\sigma_0^{-2}, \sigma^{-2}) > \delta_n \cap S_{M_n,\alpha}|\mathbf{Y}^n, \mathbf{X}^n) \to 0 \text{ in } F_0^\infty \text{ probability}, \tag{13}$$

$$\sqrt{n} \cdot \Pi(S_{M_n,\alpha}^c|\mathbf{Y}^n, \mathbf{X}^n) \to 0 \quad \text{in } F_0^\infty \text{probability}, \tag{14}$$

$$\sup_{\{\sigma \in S_{M_n,\alpha} : d_2(\sigma_0^{-2}, \sigma^{-2})^2 \leq \delta_n^2\}} \left| \sum_i X_{ij} \epsilon_i (\sigma_0^{-2}(X_i) - \sigma^{-2}(X_i))/\sqrt{n} \right| \to 0 \text{ in } F_0^\infty \text{ outer probability}. \tag{15}$$

Lemmas 7–9 show that conditions (13)–(15) hold under the assumptions of the theorem. Finally, let us consider $\int d_2(\sigma_0^{-2}, \sigma^{-2}) d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n)$. Since $d_2(\sigma_0^{-2}, \sigma^{-2}) \leq \underline{\sigma}^{-2}$,

$$\begin{aligned} & F_0^\infty\left[\int d_2(\sigma_0^{-2}, \sigma^{-2}) d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n) > \epsilon\right] \\ &\leq F_0^\infty\left[\underline{\sigma}^{-2}\Pi(d_2(\sigma_0^{-2}, \sigma^{-2}) > \epsilon/2|\mathbf{Y}^n, \mathbf{X}^n) + \epsilon/2 > \epsilon\right] \\ &= F_0^\infty\left[\Pi(d_2(\sigma_0^{-2}, \sigma^{-2}) > \epsilon/2|\mathbf{Y}^n, \mathbf{X}^n) > \epsilon/(2\underline{\sigma}^{-2})\right], \end{aligned} \tag{16}$$

[5] For standard notions and definitions used in empirical processes theory, such as outer probability and metric and bracketing entropy, see, for example, van der Vaart and Wellner (1996).

which converges to zero by conditions (13) and (14). This completes the proof of the theorem when $\hat{\beta}_{GLS}$ is used for centering.

To see that the theorem also holds when the posterior mean is used for centering, note that

$$\int \beta d\Pi(\beta|\mathbf{Y}^n, \mathbf{X}^n) = \int \overline{\beta} d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n),$$

$$\sqrt{n}\|\hat{\beta}_{GLS} - \int \overline{\beta} d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n)\|_2$$

$$\leq \int \sqrt{n}\|\hat{\beta}_{GLS} - \overline{\beta}\|_2 d\Pi(\sigma|\mathbf{Y}^n, \mathbf{X}^n)$$

and that $\sqrt{n}\|\hat{\beta}_{GLS} - \overline{\beta}\|_2$ is bounded in Lemma 6. Thus, under the posterior mean centering, the term analogous to C_n above is handled by adding and subtracting $\hat{\beta}_{GLS}$ to/from the posterior mean, and applying the triangle inequality and the proof of Lemma 6. The use of the posterior mean instead of $\hat{\beta}_{GLS}$ in D_n and E_n does not require any changes in the proof of Lemma 3. The rest of the proof is not affected by the centering. □

Lemma 2. *For two distributions P_1 and P_2 with densities p_1 and p_2 w.r.t. a measure μ, the total variation distance between P_2 truncated to a set E and P_1 can be bounded as follows*

$$\int \left| p_1 - \frac{1_E p_2}{P_2(E)} \right| d\mu \leq P_1(E^c) + \frac{P_2(E^c)}{P_2(E)} + \frac{\int |p_1 - p_2| d\mu}{P_2(E)}.$$

Proof.

$$\int |p_1 - 1_E p_2/P_2(E)| = \int_E |p_1 P_2(E) - p_2|/P_2(E) + P_1(E^c)$$

$$\leq \int_E |p_1(P_2(E) - 1) + p_1 - p_2|/P_2(E) + P_1(E^c)$$

$$\leq P_1(E^c) + (1 - P_2(E))/P_2(E) + \int |p_1 - p_2|/P_2(E). \quad \square$$

Lemma 3. *Under the assumptions of Theorem 1, E_n and D_n defined in its proof have lower bounds that do not depend on σ and converge to 1 in F_0^∞ probability.*

Proof.

$$1 - \int_{\sqrt{n}([-B,B]^d - \hat{\beta}_{GLS})} \phi z, (\sqrt{n}(\overline{\beta} - \hat{\beta}_{GLS}), (H/n)^{-1}) dz$$

$$= 1 - \int_{\sqrt{n}([-B,B]^d - \overline{\beta})} \phi(z, 0, (H/n)^{-1}) dz$$

$$\leq \sum_{i=1}^{d} \int_{z_i \notin \sqrt{n}([-B,B] - \overline{\beta}_i)} \phi(z, 0, (H/n)^{-1}) dz.$$

Next, note that

$$z'(H/n)z \geq z' \left[\left(\underline{H} + \sum_i X_i X_i'/\overline{\sigma}^2 \right) \Big/ n \right] z \geq z' z e_m^n, \tag{17}$$

where e_m^n is the smallest eigenvalue of $(\underline{H} + \sum_i X_i X_i'/\overline{\sigma}^2)/n$. As in the proof of Lemma 5,

$$|H/n| \leq \left| \left(\underline{H} + \sum_i X_i X_i'/\underline{\sigma}^2 \right) \Big/ n \right|. \tag{18}$$

Using the bound on $\|(H/n)^{-1}\|_2$ from Lemma 6, we get

$$|\overline{\beta}_i| \leq \|\overline{\beta}\|_2 \leq \|(H/n)^{-1}\|_2 \cdot \left\| \left(\underline{H}\underline{\beta} + \sum_i |X_i Y_i|/\underline{\sigma}^2 \right) \Big/ n \right\|_2$$

$$\leq \frac{\overline{\sigma}^2 \left\| \left(\underline{H}\underline{\beta} + \sum_i |X_i Y_i|/\underline{\sigma}^2 \right) \Big/ n \right\|_2}{e_{\min}\left(\left(\underline{H}\overline{\sigma} + \sum_i X_i X_i' \right) \Big/ n \right)}$$

$$= F_n \xrightarrow{F_0^\infty} F = \frac{\overline{\sigma}^2 \|E|X_i Y_i|/\underline{\sigma}^2\|_2}{e_{\min}(E(X_i X_i'))}. \tag{19}$$

Note that by Assumption 4 of Theorem 1, $B > F$. From (17)–(19),

$$\int_{z_i \notin \sqrt{n}([-B,B] - \overline{\beta}_i)} \phi(z, 0, (H/n)^{-1}) dz$$

$$\leq 2 \int_{z_i \geq \sqrt{n}(B - |\overline{\beta}_i|)} \phi(z, 0, (H/n)^{-1}) dz$$

$$\leq 2 \left| \left(\underline{H} + \sum_i X_i X_i'/\underline{\sigma}^2 \right) \Big/ n \right|^{0.5}$$

$$\times \int_{z_i \geq \sqrt{n}(B - F_n)} \exp\{-0.5 z' z e_m^n\} (2\pi)^{-d/2} dz$$

$$\leq 2 \left| \left(\underline{H} + \sum_i X_i X_i'/\underline{\sigma}^2 \right) \Big/ n \right|^{0.5} (e_m^n)^{-d/2}$$

$$\times \int_{z_i \geq \sqrt{n}(B - F_n) e_m^n} \exp\{-0.5 z_i^2\} (2\pi)^{-1/2} dz.$$

For $z \geq 1$, the normal CDF can be bounded as follows, $1 - \Phi(z) \leq \exp(-z^2)$. Thus, the integral in the last display is bounded by

$$\exp\{-n(B - F_n)^2 (e_m^n)^2\} + 1\{\sqrt{n}(B - F_n) e_m^n < 1\} \xrightarrow{F_0^\infty} 0,$$

where the convergence in probability follows from the convergence of F_n and e_m^n. This completes the proof for D_n. The proof of the analogous result for E_n is similar. □

Lemma 4. *Expression B_n from the proof of Theorem 1 can be bounded above by $B_n^1 + B_n^2 d_2(\sigma^{-2}, \sigma_0^{-2})$, where $B_n^1 \xrightarrow{F_0^\infty} 0$ and $B_n^2 \xrightarrow{F_0^\infty} B^2$, B^2 is a constant, and (B_n^1, B_n^2) do not depend on σ.*

Proof.

$$\|H/n - E[X_i X_i' \sigma_0(X_i)^{-2}]\|_\infty \leq \|\underline{H}/n\|_\infty$$

$$+ \sup_{\sigma \in S} \left\| \frac{1}{n} \sum_i \frac{X_i X_i'}{\sigma^2(X_i)} - E\left(\frac{X_i X_i'}{\sigma^2(X_i)} \right) \right\|_\infty$$

$$+ \left\| E\left(X_i X_i' \left(\frac{1}{\sigma^2(X_i)} - \frac{1}{\sigma_0^2(X_i)} \right) \right) \right\|_\infty.$$

The first term on the right hand side converges to zero. The second term converges to zero in outer probability by the assumed F_0-Glivenko–Cantelli class for $X_i X_i' \sigma^{-2}(X_i)$, $\sigma \in S$. By the Cauchy–Schwarz inequality and the finiteness of the fourth moments of X_i, the last term is bounded by a constant multiple of $d_2(\sigma^{-2}, \sigma_0^{-2})$. □

Lemma 5. *Expression A_n from the proof of Theorem 1 is bounded above by $A_n^1 + A_n^2 d_2(\sigma^{-2}, \sigma_0^{-2})$, where $A_n^1 \xrightarrow{F_0^\infty} 0$ and $A_n^2 \xrightarrow{F_0^\infty} A^2$, A^2 is a constant, and (A_n^1, A_n^2) do not depend on σ.*

Proof. It follows by the definition of the determinant and induction that for two $d \times d$ matrices A and B, $\big||A| - |B|\big| \leq d! \cdot d \max(\|A\|_\infty, \|B\|_\infty)^{d-1} \cdot \|A - B\|_\infty$. Thus, the numerator of A_n is bounded by a multiple of the bound on B_n derived in Lemma 4 times $\max(\|H/n\|_\infty, \|E[X_i X_i' \sigma_0(X_i)^{-2}]\|_\infty)^{d-1}$. Since $\|H/n\|_\infty \leq \|\underline{H}/n\|_\infty + \|\sum_i X_i X_i'/n\|_\infty/\underline{\sigma}^2$, the numerator of A_n is bounded

above as desired. To bound the denominator of A_n below note that for symmetric positive semidefinite matrices A and B, $A \geq B$ implies $|A| \geq |B|$ (see, for example, Lemma 1.4 in Zi-Zong (2009)). Thus, $|H/n| \geq |\sum_i X_i X_i'/n|/\overline{\sigma}^{2k}$. Since $|\sum_i X_i X_i'/n| \xrightarrow{F_0^\infty} |E[X_i X_i']| > 0$, the claim of the lemma follows. □

Lemma 6. *The following inequality holds for C_n defined in the proof of Theorem 1*

$$\sqrt{C_n} \leq C_n^1 + C_n^2 \left\| \frac{1}{\sqrt{n}} \sum_i X_i \epsilon_i \left(\frac{1}{\sigma_0(X_i)^2} - \frac{1}{\sigma(X_i)^2} \right) \right\|_2 + C_n^3 d_2(\sigma_0^{-2}, \sigma^{-2}),$$

where (C_n^1, C_n^2, C_n^3) do not depend on σ, $C_n^1 \xrightarrow{F_0^\infty} 0$, C_n^2 converges in F_0^∞ probability to a constant, and C_n^3 converges weakly to a random variable.

Proof. Plugging $Y_i = X_i'\beta_0 + \epsilon_i$ into the definition of C_n results in

$$\sqrt{C_n/\|H/n\|_2} = \left\| (H/n)^{-1} \underline{H}(\beta_0 - \underline{\beta})/\sqrt{n} + \left(\frac{1}{n} \sum_i \frac{X_i X_i'}{\sigma_0(X_i)^2} \right)^{-1} \frac{1}{\sqrt{n}} \sum_i \frac{X_i \epsilon_i}{\sigma_0(X_i)^2} - (H/n)^{-1} \frac{1}{\sqrt{n}} \sum_i \frac{X_i \epsilon_i}{\sigma(X_i)^2} \right\|_2. \quad (20)$$

The first expression on the right hand side of (20) converges to zero in probability because $\|(H/n)^{-1}\|_2$ is bounded above by a sequence converging in probability as it is shown below (see (23)). The norm of the second expression can be bounded by[6]

$$\|(H/n)^{-1}\|_2 \cdot \left\| \frac{1}{\sqrt{n}} \sum_i X_i \epsilon_i \left(\frac{1}{\sigma_0(X_i)^2} - \frac{1}{\sigma(X_i)^2} \right) \right\|_2 + \left\| \left(\frac{1}{n} \sum_i \frac{X_i X_i'}{\sigma_0(X_i)^2} \right)^{-1} - (H/n)^{-1} \right\|_2 \cdot \left\| \frac{1}{\sqrt{n}} \sum_i \frac{X_i \epsilon_i}{\sigma_0(X_i)^2} \right\|_2. \quad (21)$$

The norm of the difference in the inverses in the second line of (21) is bounded by[7]

$$\left\| \left(\frac{1}{n} \sum_i \frac{X_i X_i'}{\sigma_0(X_i)^2} \right)^{-1} \right\|_2 \cdot \|(H/n)^{-1}\|_2 \cdot \left\| \left(\sum_i X_i X_i' (\sigma_0(X_i)^{-2} - \sigma(X_i)^{-2}) - \underline{H} \right) \Big/ n \right\|_2. \quad (22)$$

Next, we separately consider the three parts of the product in (22). The first part converges to $\| \left(E(X_i X_i' \sigma_0(X_i)^{-2}) \right)^{-1} \|_2$ in probability. The second part,

$$\|(H/n)^{-1}\|_2 = \sup_x \frac{\|(H/n)^{-1}x\|_2}{\|x\|_2} = \sup_y \frac{\|(H/n)^{-1}(H/n)y\|_2}{\|(H/n)y\|_2}$$

$$= \left(\inf_y \frac{\|(H/n)y\|_2}{\|y\|_2} \right)^{-1} = \left(\inf_y \frac{\|y\|_2 \cdot \|(H/n)y\|_2}{\|y\|_2^2} \right)^{-1} \leq \left(\inf_y \frac{|y'(H/n)y|}{\|y\|_2^2} \right)^{-1} \leq \left(\inf_y \frac{|y'\left(\left(\underline{H}\overline{\sigma}^2 + \sum_i X_i X_i' \right) \Big/ n \right) y|/\overline{\sigma}^2}{\|y\|_2^2} \right)^{-1}$$

$$= \frac{\overline{\sigma}^2}{e_{\min}\left(\left(\underline{H}\overline{\sigma}^2 + \sum_i X_i X_i' \right)/n \right)} \xrightarrow{F_0^\infty} \frac{\overline{\sigma}^2}{e_{\min}(E(X_i X_i'))}, \quad (23)$$

where $e_{\min}(\cdot)$ stands for the smallest eigenvalue. In the preceding display, the first inequality on the third line follows by the Cauchy–Schwarz inequality, the second inequality follows by the positive semidefiniteness of $X_i X_i'$, and the last equality follows from the eigenvalue decomposition for symmetric matrices.[8]

The third part of the product in (22) is bounded above by

$$\left\| \frac{\underline{H}}{n} \right\| + \left\| \frac{1}{n} \sum_i \frac{X_i X_i'}{\sigma_0^2(X_i)} - E\left(\frac{X_i X_i'}{\sigma_0^2(X_i)} \right) \right\|_2 + \sup_{\sigma \in S} \left\| E\left(\frac{X_i X_i'}{\sigma^2(X_i)} \right) - \frac{1}{n} \sum_i \frac{X_i X_i'}{\sigma^2(X_i)} \right\|_2 + \left\| E\left(X_i X_i' \left(\frac{1}{\sigma^2(X_i)} - \frac{1}{\sigma_0^2(X_i)} \right) \right) \right\|_2,$$

which can be bounded as in Lemma 4 ($\|A\|_2 \leq \dim(A)\|A\|_\infty$). The bounds derived above and the Slutsky theorem imply the claim of the lemma. □

Lemma 7. *Under the assumptions of Theorem 1, (14) holds.*

Proof. Note that $\sqrt{n}\Pi(S^c_{M_n,\alpha}|\mathbf{Y}^n, \mathbf{X}^n)$ is bounded above by

$$\sqrt{n}\Pi(S^c_{M_n,\alpha}) \exp\left\{ n \left[\sup_{\beta \in [-B,B]^d, \sigma \in S} \frac{1}{n} \sum_i \log p_{\beta,\sigma}(Y_i|X_i) - \inf_{\beta \in [-B,B]^d, \sigma \in S} \frac{1}{n} \sum_i \log p_{\beta,\sigma}(Y_i|X_i) \right] \right\}.$$

Since $\log p_{\beta,\sigma}(Y_i|X_i) = -\log(\sqrt{2\pi}\sigma(X_i)) - (Y_i^2 - 2\beta' X_i Y_i + \beta' X_i X_i' \beta)/(2\sigma^2(X_i))$, $\beta \in [-B,B]^d$, and $0 < \underline{\sigma} \leq \sigma \leq \overline{\sigma} < \infty$, the expression in the square brackets is bounded above by a sufficiently large constant a.s. F_0^∞ by the strong law of large numbers. This constant can be increased to some $A > 0$ so that $\sqrt{n}\Pi(S^c_{M_n,\alpha}|\mathbf{Y}^n, \mathbf{X}^n) \leq \Pi(S^c_{M_n,\alpha}) \exp\{An\}$ for all sufficiently large n a.s. F_0^∞. Together with Assumption 3 in Theorem 1, this implies the claim of the lemma. □

Lemma 8. *Under the assumptions of Theorem 1, (13) holds.*

Proof. Let $Z_n = \sup_{\beta \in [-B,B]^d, \sigma \in S_{M_n,\alpha}} \left| \frac{1}{n} \sum_i \log \frac{p_{\beta_0,\sigma_0}(Y_i|X_i)}{p_{\beta,\sigma}(Y_i|X_i)} - E \log \frac{p_{\beta_0,\sigma_0}(Y_i|X_i)}{p_{\beta,\sigma}(Y_i|X_i)} \right|$. By Lemma 10, $d_2(\sigma_0^{-2}, \sigma^{-2})^2 C \leq E \log p_{\beta_0,\sigma_0}/p_{\beta,\sigma}$, where C is a constant in $(0, \infty)$. Thus,

$$\Pi\left(S_{M_n,\alpha} \cap d_2(\sigma_0^{-2}, \sigma^{-2})^2 > \delta_n^2 | \mathbf{Y}^n, \mathbf{X}^n \right) \leq \Pi\left(S_{M_n,\alpha} \cap E \log \frac{p_{\beta_0,\sigma_0}}{p_{\beta,\sigma}} > C\delta_n^2 \Big| \mathbf{Y}^n, \mathbf{X}^n \right)$$

[6] $\|A^{-1}a - B^{-1}b\| \leq \|A^{-1}(a-b)\| + \|(A^{-1} - B^{-1})b\| \leq \|A^{-1}\| \, \|a-b\| + \|A^{-1} - B^{-1}\| \, \|b\|$.

[7] $\|A^{-1} - B^{-1}\| = \|A^{-1}(A-B)B^{-1}\| \leq \|A^{-1}\| \, \|A-B\| \, \|B^{-1}\|$.

[8] $A = Q\Lambda Q'$, $QQ' = I$ and Λ is a diagonal matrix with eigenvalues of A on the diagonal.

$$\leq \frac{\int_{S_{M_n,\alpha}\cap E\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma})>C\delta_n^2} \exp\{n[C\delta_n^2/2 - E\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma}) + Z_n]\}d\Pi}{\int_{S_{M_n,\alpha}\cap E\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma})\leq C\delta_n^2/2} \exp\{n[C\delta_n^2/2 - E\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma}) - Z_n]\}d\Pi}$$

$$\leq \frac{\exp\{n[-C\delta_n^2/2 + 2Z_n]\}}{\Pi\left(S_{M_n,\alpha}\cap E\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma})\leq C\delta_n^2/2\right)}$$

$$\leq \exp\{-n[(1-r)C\delta_n^2/2 - 2Z_n]\},$$

where the last inequality follows from Assumptions 3 and 5 of the theorem. For $\epsilon > 0$,

$$F_0^\infty\left[S_{M_n,\alpha}\cap\sqrt{n}\Pi(d_2(\sigma_0^{-2},\sigma^{-2})^2 > \delta_n^2|\mathbf{Y}^n,\mathbf{X}^n) > \epsilon\right]$$
$$\leq F_0^{\infty *}\left[Z_n > \frac{(1-r)C\delta_n^2}{4} - \frac{\log[\sqrt{n}/\epsilon]}{2n}\right]$$
$$\leq F_0^{\infty *}\left[Z_n > \frac{(1-r)C\delta_n^2}{8}\right], \tag{24}$$

where the last inequality holds for all sufficiently large n by $\delta_n^4 n\to\infty$ (Assumption 2 of the theorem). It remains to prove that the last bound in (24) converges to zero. Since

$$\log\frac{p_{\beta_0,\sigma_0}(Y_i|X_i)}{p_{\beta,\sigma}(Y_i|X_i)}$$
$$= \frac{1}{2}\left(\log\sigma^2(X_i) - \log\sigma_0^2(X_i) - \frac{\epsilon_i^2}{\sigma_0^2(X_i)} + \frac{\epsilon_i^2}{\sigma^2(X_i)}\right.$$
$$\left. + 2\frac{\epsilon_i X_i'}{\sigma^2(X_i)}(\beta_0-\beta) + (\beta-\beta_0)'\left(\frac{X_iX_i'}{\sigma^2(X_i)}\right)(\beta-\beta_0)\right) \tag{25}$$

and $\beta,\beta_0\in[-B,B]^d$, it suffices to prove that

$$F_0^{\infty *}\left(\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_i f(\epsilon_i,X_i) - Ef(\epsilon_i,X_i)\right| \geq \tilde{C}\delta_n^2\right)\to 0,$$
$$\forall\,\tilde{C}>0, \tag{26}$$

for the following four classes $\mathcal{F}_n$: $\{\log\sigma(X_i),\ \sigma\in S_{M_n,\alpha}\}$, $\{\epsilon_i^2/\sigma(X_i)^2,\ \sigma\in S_{M_n,\alpha}\}$, $\{\epsilon_i X_{ik}/\sigma(X_i)^2,\ \sigma\in S_{M_n,\alpha}\}$, and $\{X_{ij}X_{ik}/\sigma(X_i)^2,\ \sigma\in S_{M_n,\alpha}\}$. Convergence at rate δ_n^2 for expressions in (25) involving σ_0 follows from the Chebyshev's inequality, the existence of second moments (Assumption 1), and $\delta_n^4 n\to\infty$.

It follows from inequalities (30) and (31) on p. 31 in Pollard (1984) that (26) is bounded above by

$$4E^*\min\left\{1, 2N_1\left(\frac{\tilde{C}\delta_n^2}{8},\mathcal{F}_n,L_1(F_{0,n})\right)\right.$$
$$\left.\times\exp\left[-\frac{1}{2}n\left(\frac{\tilde{C}\delta_n^2}{8}\right)^2\frac{1}{\max_{j=1,\ldots,N_1}\sum_i g_j^2(\epsilon_i,X_i)}\right]\right\},$$

where $N_1 = N_1(\cdot,\mathcal{F}_n,L_1(F_{0,n}))$ is the covering number w.r.t. $L_1(F_{0,n})$ distance, $F_{0,n}$ is the empirical measure corresponding to F_0, and g_j's are the centers of the covering balls. Lemma 2.3.7 in van der Vaart and Wellner (1996) on p. 112 gives a proof of Eq. (30) in Pollard (1984) using outer probabilities and expectations to avoid measurability problems.

As $\mathcal{X}$ is assumed to be bounded and convex, it follows from lemma 2.7.1 in van der Vaart and Wellner (1996) that the metric entropy in the sup norm of $S_{M,\alpha}$ is bounded by

$$\log N(\epsilon,S_{M,\alpha},\|\cdot\|_\infty)\leq K(M/\epsilon)^{d/\alpha}, \tag{27}$$

where $K>0$ is a constant that does not depend on (M,ϵ). For $\mathcal{F}_n = \{\epsilon_i^2/\sigma(X_i)^2,\ \sigma\in S_{M_n,\alpha}\}$,

$$1/n\sum_i|f_1(\epsilon_i,X_i) - f_2(\epsilon_i,X_i)| \leq \sup|\sigma_1^{-2}-\sigma_2^{-2}|\sum_i\epsilon_i^2/n.$$

Then, the bound on the uniform metric entropy for $S_{M_n,\alpha}$ in (27) implies

$$N_1\left(\epsilon,\mathcal{F}_n,L_1(F_{0,n})\right)\leq\exp\{C_1(Q_nM_n/\epsilon)^{d/\alpha}\},$$

where C_1 is a constant and $Q_n = \sum_i\epsilon_i^2/n$ converges a.s. to a constant. Also note that

$$\max_{j=1,N_1}\sum_i g_j^2(\epsilon_i,X_i)\leq\underline{\sigma}^{-4}R_n,$$

where $R_n = \sum_i\epsilon_i^4/n$ converges a.s. to a constant. Thus, (26) is bounded above by

$$4E\min\left\{1, 2\exp\left[C_2Q_n^{d/\alpha}M_n^{d/\alpha}\delta_n^{-2d/\alpha} - C_3n\delta_n^4/R_n\right]\right\},$$

where C_2 and C_3 are positive constants. Since the integrand is uniformly bounded by 1, the limsup version of Fatou's lemma implies that the previous display converges to zero under Assumption 2 of Theorem 1. Claim (26) for the other three classes $\mathcal{F}_n$ follow by the same argument with modified R_n and Q_n. □

Lemma 9. *Under the assumptions of Theorem 1, (15) holds.*

Proof. The proof is based on lemma 19.34 in van der Vaart (1998). Define

$$\mathcal{F} = \{X_{ij}\epsilon_i[\sigma_0^{-2}(X_i) - \sigma^{-2}(X_i)]:\ d_2(\sigma_0^{-2},\sigma^{-2})^2\leq\delta_n^2,\sigma\in S_{M_n,\alpha}\}.$$

For any $f\in\mathcal{F}$, $Ef^2\leq C_1\delta_n^2 = \delta^2$, where C_1 is a positive constant and δ is defined by the last equality. For $f_1,f_2\in\mathcal{F}$, $E(f_1-f_2)^2\leq C_2\sup|\sigma_1-\sigma_2|$, where C_2 is a positive constant. Thus, by (27), the bracketing $L_2(F_0)$ entropy $\log N_{[]}(\delta,\mathcal{F},L_2(F_0))\leq C_3[M_n/\delta]^{d/\alpha}$ and

$$a(\delta) = \delta/\sqrt{\log N_{[]}(\delta,\mathcal{F},L_2(F_0))}\geq C_4\delta_n^{(1+d/(2\alpha))}M_n^{-d/(2\alpha)}.$$

A bound on the bracketing integral is proportional to $\delta^{1-d/(2\alpha)}M_n^{d/(2\alpha)}$ and

$$J_{[]}(\delta,\mathcal{F},L_2(F_0))\leq C_5\delta_n^{(1-d/(2\alpha))}M_n^{d/(2\alpha)}.$$

By lemma 19.34 in van der Vaart (1998),

$$E^*\sup_{f\in\mathcal{F}}\left|\sum_i f(\epsilon_i,X_i)/\sqrt{n}\right| \leq J_{[]}(\delta,\mathcal{F},L_2(F_0))$$
$$+\sqrt{n}E(|X_{ij}\epsilon_i|1\{|X_{ij}\epsilon_i|>\sqrt{n}a(\delta)\}).$$

By Assumption 2 of Theorem 1, this display converges to zero. Thus, the claim of the lemma is proved. □

Lemma 10. *For some positive constants $\underline{C}$ and $\bar{C}$*

$$E(\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma}))\geq\underline{C}[\|\beta-\beta_0\|_2^2 + E(\sigma_0^2-\sigma^2)^2] \tag{28}$$
$$E(\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma}))\leq\bar{C}[\|\beta-\beta_0\|_2^2 + E(\sigma_0^2-\sigma^2)^2]. \tag{29}$$

Proof. The law of iterated expectations implies

$$E\left(\log\frac{p_{\beta_0,\sigma_0}}{p_{\beta,\sigma}}\right) = \frac{1}{2}E\left(\log\frac{\sigma^2(X_i)}{\sigma_0^2(X_i)} + \frac{\sigma_0^2(X_i)-\sigma^2(X_i)}{\sigma^2(X_i)}\right.$$
$$\left.+(\beta-\beta_0)'\left(\frac{X_iX_i'}{\sigma^2(X_i)}\right)(\beta-\beta_0)\right).$$

First, for $e_{\min}$ and $e_{\max}$ denoting the smallest and largest eigenvalues, note that

$$\frac{e_{\min}(E(X_iX_i'))}{\bar{\sigma}^2}\|\beta-\beta_0\|_2^2\leq(\beta-\beta_0)'E\left(\frac{X_iX_i'}{\sigma^2(X_i)}\right)(\beta-\beta_0)$$
$$\leq\frac{e_{\max}(E(X_iX_i'))}{\underline{\sigma}^2}\|\beta-\beta_0\|_2^2.$$

Second, let $\sigma_0^2/\sigma^2 = z$ and $q(z) = (z-1-\log z)/(z-1)^2$. Note that $q(z)$ is well defined, positive, and monotonically decreasing

on $(0,\infty)$. Thus, for any $z \in [\underline{\sigma}^2/\overline{\sigma}^2, \overline{\sigma}^2/\underline{\sigma}^2]$, $q(\overline{\sigma}^2/\underline{\sigma}^2) \le q(z) \le q(\underline{\sigma}^2/\overline{\sigma}^2)$. From this inequality,

$$\frac{E(\sigma_0^2-\sigma^2)^2}{\overline{\sigma}^4} q(\overline{\sigma}^2/\underline{\sigma}^2) \le E\left(\log\frac{\sigma^2}{\sigma_0^2} + \frac{\sigma_0^2-\sigma^2}{\sigma^2}\right) \le \frac{E(\sigma_0^2-\sigma^2)^2}{\underline{\sigma}^4} q(\underline{\sigma}^2/\overline{\sigma}^2). \tag{30}$$

Thus, inequalities (28) and (29) are proved. □

Proof. Proposition 1.

First, consider Assumption 5. By Lemma 10 and $E(\sigma_0^2-\sigma^2)^2 \le [2\overline{\sigma}\sup|\sigma-\sigma_0|]^2$,

$$\Pi\left(E\log\frac{p_{\beta_0,\sigma_0}}{p_{\beta,\sigma}} \le C\delta_n^2\right) \ge \Pi\left(\|\beta-\beta_0\|_2^2 \le \frac{C\delta_n^2}{2\bar{C}}\right) \times \Pi\left(\sup|\sigma-\sigma_0| \le \frac{C^{1/2}\delta_n}{2\overline{\sigma}(2\bar{C})^{1/2}}\right). \tag{31}$$

A lower bound on the first term on the right hand side is proportional to δ_n^d, which is polynomial in n and, thus, can be ignored.

Since $\underline{\sigma} < \sigma_0 < \overline{\sigma}$, for all sufficiently small ε_n

$$\Pi\left(\sup_x|\sigma(x)-\sigma_0(x)| \le 2\varepsilon_n\right) = \frac{\mathcal{W}\left(\sup_x\left|(I^jW)(x)+\sum_{j=0}^J Z_j x^j/j! - \sigma_0(x)\right| \le 2\varepsilon_n\right)}{\mathcal{W}(T)}.$$

Theorems 2.1 and 4.1 in van der Vaart and van Zanten (2008) imply the following lower bound for prior concentration probability around σ_0,

$$\Pi\left(\sup_x|\sigma(x)-\sigma_0(x)| \le 2\varepsilon_n\right) \ge \mathcal{W}(T)^{-1}\exp\{-n\varepsilon_n^2\}, \quad \varepsilon_n \propto n^{-\alpha_0/(2J+2)}$$

when $\alpha_0 \le J+1/2$. Thus, Assumption 5 of Theorem 1 holds as long as

$$n^{-\alpha_0/(2J+2)}/\delta_n \to 0, \quad \alpha_0 \le J+1/2. \tag{32}$$

Next, let us consider Assumption 3 of Theorem 1 with $\alpha = J$.

$$\left[(I^jW)(x)+\sum_{j=0}^J Z_j x^j/j! \in S^c_{M_n,\alpha}\right] \subset \left[\sup_x|W(x)| \ge M_n/4\right] \bigcup \left[\exists j, |Z_j| \ge M_n/(2(J+1))\right].$$

Thus, $\Pi(S^c_{M_n,\alpha}) \le \mathcal{W}(T)^{-1}\cdot\left(\mathcal{W}[\sup_x|W(x)| \ge M_n/4] + (J+1)\mathcal{W}[|z_j| \ge M_n/(2(J+1))]\right)$. Note that $\mathcal{W}(\sup_x|W(x)| \ge M_n/4) \le 4[1-\Phi(M_n/4)]$ by the standard results on barrier crossing probabilities for W, see, for example, theorem 7.1 on p.314 in Shorack (2000). Then, Assumption 3 of Theorem 1 holds if

$$M_n \ge n^{(1+r_1)/2}, \quad \text{for some } r_1 > 0. \tag{33}$$

Thus, to prove the proposition it suffices to find δ_n and M_n that satisfy conditions (3), (32) and (33). These conditions are satisfied if

$$\gamma_2 < \frac{\alpha_0}{2J+2}, \quad \gamma_2 < \frac{1}{6+J^{-1}}, \quad \alpha_0 \le J+1/2, \quad \text{and} \quad \gamma_1 = \frac{1+r_1}{2} < (2J-1)\gamma_2.$$

Values of (γ_1,γ_2,r_1) satisfying these inequalities can be chosen as long as

$$\frac{1}{2(2J-1)} < \min\left\{\frac{1}{6+J^{-1}}, \frac{\alpha_0}{2J+2}\right\},$$

which holds for any $J \ge 3$ and $\alpha_0 > (J+1)/(2J-1)$. □

Proof. (Proposition 2)

For $\alpha = \lfloor d/2\rfloor + 1$, γ_1 from (9), γ_3 from (7), $\delta_n = n^{-\gamma_2}$, and $\gamma_2 = \frac{1-s}{d[6+d/(\lfloor d/2\rfloor+1)]}$, it can be verified by a direct calculation that inequalities in (3) hold. Thus, Assumption 2 of Theorem 1 holds.

Simple algebraic manipulations deliver the following convergence results that are used below,

$$\frac{M_n}{m_1^{\lfloor d/2\rfloor+1/2}} \to 0, \qquad \delta_n m_1^{1/2} \to \infty,$$
$$\frac{m_1^d \log m_1}{n\delta_n^2} \to 0 \quad \text{as } n\to\infty. \tag{34}$$

By Part (iv) of Lemma 11, the assumed existence and uniform boundedness of partial derivatives up to order $\lfloor d/2\rfloor+2$ for σ_0 implies that for some $N \in \mathbb{N}$ and a constant $M > 0$

$$|D^{m,\lambda}\sigma_0(x)| \le M \tag{35}$$

for all $x \in [0,1]^d$, all $m_i \ge N$, $i = 1,\dots,d$ and all λ with $|\lambda| = \sum_{j=1}^d \lambda_j \le \lfloor d/2\rfloor + 1$. Constant M can be chosen sufficiently large so that it is also a Lipschitz constant for σ_0: $|\sigma_0(x_1)-\sigma_0(x_2)| \le M\|x_1-x_2\|_2$.

First, let us verify that Π_n puts probability 1 on $S_{M_n,\lfloor d/2\rfloor+1}$. For any $x \in [0,1]^d$, $\sum_{j_i=0}^{m_i}\binom{m_i}{j_i}x_i^{j_i}(1-x_i)^{m_i-j_i} = (x_i+1-x_i)_i^m = 1$ and

$$\sum_{o\le j\le m}\prod_{i=1}^d\binom{m_i}{j_i}x_i^{j_i}(1-x_i)^{m_i-j_i} = 1.$$

Thus, for g^m satisfying (8),

$$\underline{\sigma} \le \min_{o\le j\le m} g_j^m \le B^m(x;g^m) \le \max_{o\le j\le m} g_j^m \le \overline{\sigma}. \tag{36}$$

Next, let us consider partial derivatives of $B^m(x;g^m)$. By (5), (6), (8) and (36)

$$|\partial^\lambda B^m(x,g^m)| = |B^{m-\lambda}(x;D^{m,\lambda}(G^{m-\lambda};g^m))| \le \max|D^{m,\lambda}(G^{m-\lambda};g^m)| \le M_n.$$

Thus, Assumption 3 of Theorem 1 holds.

Next, consider Assumption 5 of Theorem 1. By (31), it suffices to show that

$$\Pi_n\left(g^m : \sup_x|B^m(x;g^m)-\sigma_0(x)| \le C^{1/2}\delta_n/(2\overline{\sigma}(2\bar{C})^{0.5})\right) \ge \exp\{-nrC\delta_n^2\}.$$

By theorem B.7 in Appendix B of Heitzinger (2002), $|B^m(x;g^{\star m})-\sigma_0(x)| \le \frac{Md^{0.5}}{2m_1^{0.5}}$ for $g^{\star m} = \sigma_0(G^m)$. For g^m satisfying $\max_j|g_j^m - g_j^{\star m}| \le (M_n-M)/((2m_1)^{\lfloor d/2\rfloor+1})$,

$$\sup_x|B^m(x;g^m)-\sigma_0(x)| \le \sup_x|B^m(x;g^{\star m})-\sigma_0(x)| + \sup_x|B^m(x;g^{\star m})-B^m(x;g^m)| \le [M/2]d^{0.5}/m_1^{0.5} + (M_n-M)/((2m_1)^{\lfloor d/2\rfloor+1}) \le Md^{0.5}/m_1^{0.5},$$

where the last inequality holds for all sufficiently large m_1 by the first convergence result in (34). Also, note that for such g^m, the prior truncation constraint (8) holds (for $\lambda \le \lfloor d/2\rfloor+1$ and $x \in G^{m-\lambda}$, $|D^{m,\lambda}(x;g^{\star m})| = |D^{m,\lambda}(\sigma_0)(x)| \le M$, and, thus, by an analog of (39)

in Lemma 11, $|D^{m,\lambda}(x;g^m)| \le |D^{m,\lambda}(x;g^m) - D^{m,\lambda}(x;g^{\star m})| + M \le M_n$). Therefore,

$$\begin{aligned}
&\Pi_n\left(\sup_x |B^m(x;g^m) - \sigma_0(x)| \le \frac{C^{1/2}\delta_n}{2\overline{\sigma}(2\bar{C})^{0.5}}\right)\\
&\ge \Pi_n\left(\sup_x |B^m(x;g^m) - \sigma_0(x)| \le \frac{Md^{0.5}}{m_1^{0.5}}\right)\\
&\ge \Pi_n\left(\max_j |g_j^m - g_j^{\star m}| \le \frac{M_n - M}{(2m_1)^{\lfloor d/2\rfloor+1}}\right)\\
&\ge \underline{\pi}_g^{(m_1+1)^d}\cdot\left(\frac{M_n - M}{(2m_1)^{\lfloor d/2\rfloor+1}}\right)^{(m_1+1)^d},
\end{aligned}$$

where the first inequality holds because $\delta_n m_1^{0.5} \to \infty$ and the last inequality follows because the prior density is assumed to be bounded from below by $\underline{\pi}_g^{(m_1+1)^d}$ on its support (8).

The last bound in the above inequality combined with $[(m(n)+1)^d \log m(n)]/(n\delta_n^2) \to 0$ implies Assumption 5 of Theorem 1. □

Lemma 11. (i) *For $x = (x_i, x_{-i}) \in [0,1]^d$, where $x_i \in [0,1]$,*

$$\begin{aligned}
(D_i^{m_i,\lambda_i} f)(x_i, x_{-i}) &= \left[\prod_{j_i=1}^{\lambda_i}(m_i - \lambda_i + j_i)\right]\sum_{j_i=0}^{\lambda_i} f\\
&\times\left(\frac{m_i-\lambda_i}{m_i}x_i + \frac{j_i}{m_i}, x_{-i}\right)\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i}. \qquad (37)
\end{aligned}$$

(ii) *For multi-indices $m \ge \lambda$,*

$$\begin{aligned}
(D^{m,\lambda}f)(x) &= \left[\prod_{i=1}^d\prod_{j_i=1}^{\lambda_i}(m_i - \lambda_i + j_i)\right]\sum_{0\le j\le\lambda} f\\
&\times n\left(\frac{m-\lambda}{m}x + \frac{j}{m}\right)\prod_{i=1}^d\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i}, \qquad (38)
\end{aligned}$$

where operations on vectors (x, λ, m, j) inside f are coordinate-wise.

(iii) *For functions f_1 and f_2 from $[0,1]^d$ to $\mathbb{R}$,*

$$|(D^{m,\lambda}f_1)(x) - (D^{m,\lambda}f_1)(x)| \le (2\max m_i)^{|\lambda|}\sup_x |f_1(x) - f_2(x)|. \qquad (39)$$

(iv) *If all components of m are the same ($m_i = m_1$) and f has bounded partial derivatives up to order $|\lambda|+1$, then*

$$\sup_x\left|(D^{m,\lambda}f)(x) - (\partial^\lambda f)\left(\frac{m-\lambda}{m}x\right)\right| \to 0 \quad \text{as } m_1 \to \infty. \qquad (40)$$

(v) *Eq. (4) holds.*

Proof. Formula (37) follows by induction on λ_i: for $\lambda_i = 1$, the formula holds by the definition of $D_i^{m_i,\lambda_i}$; assume it holds for λ_i. Then $D_i^{m_i,\lambda_i+1}f(x)/\prod_{j_i=1}^{\lambda_i+1}(m_i - \lambda_i + j_i)$ is equal to

$$\begin{aligned}
&\sum_{j_i=0}^{\lambda_i} f\left(\frac{m_i-\lambda_i}{m_i}\left[x_i\left(1 - \frac{1}{m_i-\lambda_i}\right) + \frac{1}{m_i-\lambda_i}\right] + \frac{j_i}{m_i}, x_{-i}\right)\\
&\quad\times\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i} - \sum_{j_i=0}^{\lambda_i} f\left(\frac{m_i-\lambda_i}{m_i}\left[x_i\left(1 - \frac{1}{m_i-\lambda_i}\right)\right]\right.\\
&\quad\left.+\frac{j_i}{m_i}, x_{-i}\right)\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i}\\
&= f\left(\frac{m_i-\lambda_i-1}{m_i}x_i + \frac{0}{m_i}, x_{-i}\right)(-1)^{\lambda_i+1}\\
&\quad + f\left(\frac{m_i-\lambda_i-1}{m_i}x_i + \frac{\lambda_i+1}{m_i}, x_{-i}\right)\\
&\quad + \sum_{j_i=1}^{\lambda_i} f\left(\frac{m_i-\lambda_i-1}{m_i}x_i + \frac{j_i}{m_i}, x_{-i}\right)\\
&\quad\times\left[(-1)^{\lambda_i-(j_i-1)}\binom{\lambda_i}{j_i-1} - (-1)^{\lambda_i - j_i}\binom{\lambda_i}{j_i}\right]
\end{aligned}$$

and (37) is proved since the last square bracket term is equal to $(-1)^{\lambda_i+1-j_i}\binom{\lambda_i+1}{j_i}$.

Formula (38) follows from repeated application of (37). Inequality (39) follows immediately from (38) and $\sum_{j_i=0}^{\lambda_i}\binom{\lambda_i}{j_i} = 2^{\lambda_i}$. To prove (40), plug the following Taylor expansion

$$\begin{aligned}
f\left(\frac{m-\lambda}{m}x + \frac{j}{m}\right) &= \left[\sum_{|l|\le|\lambda|}\partial^l f\left(\frac{m-\lambda}{m}x\right)\frac{1}{l!}\prod_{i=1}^d\left(\frac{j_i}{m_i}\right)^{l_i}\right]\\
&\quad + R(x,\lambda,m,j)\cdot(1/m_1)^{|\lambda|+1},
\end{aligned}$$

where $R(x,\lambda,m,j)$ is bounded by a constant independent of (x,λ,m,j) and $l! = l_1!\cdots l_k!$, into (38) to obtain

$$\begin{aligned}
&\frac{(D^{m,\lambda}f)(x)}{\prod_{i=1}^d\prod_{j_i=1}^{\lambda_i}(m_1 - \lambda_i + j_i)}\\
&= \sum_{|l|\le|\lambda|}\left[\partial^l f\left(\frac{m-\lambda}{m}x\right)\frac{1}{l!}\cdot\prod_{i=1}^d\sum_{j_i=0}^{\lambda_i}\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i}\left(\frac{j_i}{m_1}\right)^{l_i}\right]\\
&\quad + \sum_{0\le j\le\lambda} R(x,\lambda,m,j)\cdot(1/m_1)^{|\lambda|+1}\prod_{i=1}^d\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i}.
\end{aligned}$$

By properties of binomial coefficients, $\sum_{j_i=0}^{\lambda_i}\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i}j_i^{l_i}$ is equal to zero if $l_i < \lambda_i$ and $\lambda_i!$ if $l_i = \lambda_i$ (Ruiz, 1996). Thus,

$$\begin{aligned}
(D^{m,\lambda}f)(x) &= \frac{\prod_{i=1}^d\prod_{j_i=1}^{\lambda_i}(m_1 - \lambda_i + j_i)}{m_1^{|\lambda|}}\cdot\left[\partial^\lambda f\left(\frac{m-\lambda}{m}x\right)\right.\\
&\quad\left.+\frac{1}{m_1}\sum_{0\le j\le\lambda} R(x,\lambda,m,j)\cdot\prod_{i=1}^d\binom{\lambda_i}{j_i}(-1)^{\lambda_i - j_i}\right]
\end{aligned}$$

and (40) follows as $\partial^\lambda f$ is uniformly bounded.

Eq. (4) follows by a direct calculation,

$$\begin{aligned}
\frac{\partial}{\partial x_i}(B_i^{m_i}f)(x) &= \sum_{j_i=1}^{m_i} f(j_i/m_i, x_{-i})\binom{m_i}{j_i}j_i x_i^{j_i-1}(1-x_i)^{m_i - j_i}\\
&\quad - \sum_{j_i=0}^{m_i-1} f(j_i/m_i, x_{-i})\binom{m_i}{j_i}(m_i - j_i)x_i^{j_i}(1-x_i)^{m_i-1-j_i}\\
&= \sum_{j_i=0}^{m_i-1} m_i f((j_i+1)/m_i, x_{-i})\binom{m_i-1}{j_i}x_i^{j_i}(1-x_i)^{m_i-1-j_i}\\
&\quad - \sum_{j_i=0}^{m_i-1} m_i f(j_i/m_i, x_{-i})\binom{m_i-1}{j_i}x_i^{j_i}(1-x_i)^{m_i-1-j_i}\\
&= \sum_{j_i=0}^{m_i-1}(D_i^{m_i}f)(j_i/(m_i-1), x_{-i})\binom{m_i-1}{j_i}x_i^{j_i}(1-x_i)^{m_i-1-j_i}\\
&= (B_i^{m_i-1}D_i^{m_i}f)(x). \quad \square
\end{aligned}$$

Acknowledgments

The author is grateful to Ulrich Mueller, Justinas Pelenis, and Chris Sims for helpful discussions. He also thanks anonymous referees for useful suggestions. This paper is based upon work partially supported by the National Science Foundation under Grant No. SES-1260861.

References

Andrews, D.W., 1994. Empirical process methods in econometrics. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics. In: Handbook of Econometrics, vol. 4. Elsevier, pp. 2247–2294 (Chapter 37).
Barron, A., 1988. The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions. University of Illinois, Dept. of Statistics.
Barron, A., Schervish, M.J., Wasserman, L., 1999. The consistency of posterior distributions in nonparametric problems. Ann. Statist. 27, 536–561.
Belitser, E., Ghosal, S., 2003. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. Ann. Statist. 31, 536–559. Dedicated to the memory of Herbert E. Robbins.
Bickel, P.J., Kleijn, B.J.K., 2012. The semiparametric Bernstein-von Mises theorem. Ann. Statist. 40, 206–237.
Burda, M., Prokhorov, A., 2013. Copula based factorization in Bayesian multivariate infinite mixture models. Working Papers, University of Toronto, Department of Economics.
Carroll, R.J., 1982. Adapting for heteroscedasticity in linear models. Ann. Statist. 10, 1224–1233.
Castillo, I., 2012. A semiparametric Bernstein-von Mises theorem for Gaussian process priors. Probab. Theory Related Fields 152, 53–99.
Castillo, I., Nickl, R., 2013. Nonparametric Bernstein-von Mises theorems in Gaussian white noise. Ann. Statist. 41, 1999–2028.
Castillo, I., Nickl, R., 2014. On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. Ann. Statist. 42, 1941–1969.
Castillo, I., Rousseau, J., 2013. A General Bernstein-von Mises theorem in semiparametric models. ArXiv:1305.4482.
Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. J. Econometrics 34, 305–334.
Chernozhukov, V., Hong, H., 2003. An MCMC approach to classical estimation. J. Econometrics 115, 293–346.
Chib, S., Greenberg, E., 2013. On conditional variance estimation in nonparametric regression. Stat. Comput. 23, 261–270.
Chung, Y., Dunson, D.B., 2009. Nonparametric bayes conditional distribution modeling with variable selection. J. Amer. Statist. Assoc. 104, 1646–1660.
De Iorio, M., Muller, P., Rosner, G.L., MacEachern, S.N., 2004. An ANOVA model for dependent random measures. J. Amer. Statist. Assoc. 99, 205–215.
Dunson, D.B., Park, J.-H., 2008. Kernel stick-breaking processes. Biometrika 95, 307–323.
Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. Wiley-Interscience.
Geweke, J., Keane, M., 2007. Smoothly mixing regressions. J. Econometrics 138, 252–290.
Ghosal, S., 2001. Convergence rates for density estimation with Bernstein polynomials. Ann. Statist. 29, 1264–1280.
Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 1999. Posterior consistency of dirichlet mixtures in density estimation. Ann. Statist. 27, 143–158.
Ghosal, S., Ghosh, J.K., Vaart, A.W.v.d., 2000. Convergence rates of posterior distributions. Ann. Statist. 28, 500–531.
Ghosal, S., Lember, J., Van Der Vaart, A., 2003. On Bayesian adaptation. In: Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics, Part II (2002), vol. 79. pp. 165–175.
Ghosal, S., Lember, J., van der Vaart, A., 2008. Nonparametric Bayesian model selection and averaging. Electron. J. Stat. 2, 63–89.
Ghosh, J., Ramamoorthi, R., 2003. Bayesian Nonparametrics, first ed. Springer.
Gine, E., Nickl, R., 2011. Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. Ann. Statist. 39, 2883–2911.
Goldberg, P.W., Williams, C.K.I., Bishop, C.M., 1998. Regression with input-dependent noise: a Gaussian process treatment. In: Advances in Neural Information Processing Systems, vol. 10. MIT Press, pp. 493–499.
Gourieroux, C., Monfort, A., Trognon, A., 1984. Pseudo maximum likelihood methods: Theory. Econometrica 52, 681–700.
Griffin, J.E., Steel, M.F.J., 2006. Order-based dependent dirichlet processes. J. Amer. Statist. Assoc. 101, 179–194.
Heitzinger, C., 2002. Simulation and Inverse Modeling of Semiconductor Manufacturing Processes, Technische Universität Wien. http://www.iue.tuwien.ac.at/phd/heitzinger/node130.html.
Huang, T.-M., 2004. Convergence rates for posterior distributions and adaptive estimation. Ann. Statist. 32, 1556–1593.
Huber, P., 1967. The behavior of the maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. University of California Press, Berkeley, pp. 221–233.
Jaynes, E.T., 1957. Information theory and statistical mechanics. Phys. Rev. 106, 620–630.
Kato, K., 2013. Quasi-Bayesian analysis of nonparametric instrumental variables models. Ann. Statist. 41, 2359–2390.
Kemperman, J.H.B., 1969. On the optimum rate of transmitting information. Ann. Math. Statist. 40, 2156–2177.
Kleijn, B., Knapik, B., 2012. Semiparametric posterior limits under local asymptotic exponentiality. ArXiv:1210.6204.
Kleijn, B., van der Vaart, A., 2006. Misspecification in infinite-dimensional Bayesian statistics. Ann. Statist. 34, 837–877.
Kleijn, B., van der Vaart, A., 2012. The Bernstein-Von-Mises theorem under misspecification. Electron. J. Stat. 6, 354–381.
Kruijer, W., Rousseau, J., van der Vaart, A., 2010. Adaptive Bayesian density estimation with location-scale mixtures. Electron. J. Stat. 4, 1225–1257.
Kruijer, W., van der Vaart, A., 2008. Posterior convergence rates for Dirichlet mixtures of beta densities. J. Statist. Plann. Inference 138, 1981–1992.
Lancaster, T., 2003. A note on bootstraps and robustness.
Liang, S., Carlin, B.P., Gelfand, A.E., 2009. Analysis of minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. Ann. Appl. Statist. 3, 943–962.
Lorentz, G.G., 1986. Bernstein Polynomials. Chelsea Pub. Co., New York, NY.
MacEachern, S.N., 1999. Dependent nonparametric processes. In: ASA Proceedings of the Section on Bayesian Statistical Science.
Majer, P., 2012. Multivariate Bernstein polynomials for approximation of derivatives. MathOverflow. http://mathoverflow.net/questions/111257 (version: 2012-11-03).
Müller, U.K., 2013. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. Econometrica 81, 1805–1849.
Norets, A., 2010. Approximation of conditional densities by smooth mixtures of regressions. Ann. Statist. 38, 1733–1766.
Norets, A., Pelenis, J., 2014. Posterior consistency in conditional density estimation by covariate dependent mixtures. Econometric Theory 30, 606–646.
Panov, M., Spokoiny, V., 2013. Finite sample Bernstein-von Mises theorem for semiparametric problems. ArXiv:1310.7796.
Pati, D., Dunson, D.B., Tokdar, S.T., 2013. Posterior consistency in conditional distribution estimation. J. Multivariate Anal. 116, 456–472.
Pelenis, J., 2014. Bayesian regression with heteroscedastic error density and parametric mean function. J. Econometrics 178 (3), 624–638.
Peng, F., Jacobs, R.A., Tanner, M.A., 1996. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. J. Amer. Statist. Assoc. 91, 953–960.
Petrone, S., 1999. Bayesian density estimation using bernstein polynomials. Canad. J. Statist. 27, 105–126.
Petrone, S., Wasserman, L., 2002. Consistency of Bernstein polynomial posteriors. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 79–100.
Poirier, D., 2011. Bayesian interpretations of heteroskedastic consistent covariance estimators using the informed bayesian bootstrap. Econometric Rev. 30, 457–468.
Pollard, D., 1984. Convergence of stochastic processes. In: Springer Series in Statistics. Springer Verlag GMBH.
Rivoirard, V., Rousseau, J., 2012. Bernstein-von Mises theorem for linear functionals of the density. Ann. Statist. 40, 1489–1523.
Robinson, P.M., 1987. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. Econometrica 55, 875–891.
Rousseau, J., 2010. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. Ann. Statist. 38, 146–180.
Rubin, D., 1981. The Bayesian bootstraps. Ann. Statist. 9, 130–134.
Ruiz, S.M., 1996. An algebraic identity leading to Wilson's theorem. Math. Gaz. 80, 579–582.
Schwartz, L., 1965. On Bayes procedures. Z. Wahrscheinlichkeitstheor. Verwandte Geb. 4, 10–26.
Scricciolo, C., 2006. Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. Ann. Statist. 34, 2897–2920.
Shen, X., 2002. Asymptotic normality of semiparametric and nonparametric posterior distributions. J. Amer. Statist. Assoc. 97, 222–235.
Shen, X., Wasserman, L., 2001. Rates of convergence of posterior distributions. Ann. Statist. 29, 687–714.
Shorack, G., 2000. Probability for Statisticians. Springer, New York.
Tokdar, S., 2007. Towards a faster implementation of density estimation with logistic Gaussian process priors. J. Comput. Graph. Statist. 16, 633–655.
Tokdar, S., Zhu, Y., Ghosh, J., 2010. Bayesian density regression with logistic Gaussian process and subspace projection. Bayesian Anal. 5, 319–344.
Tokdar, S.T., Ghosh, J.K., 2007. Posterior consistency of logistic Gaussian process priors in density estimation. J. Statist. Plann. Inference 137, 34–42.
van der Vaart, A., 1998. Asymptotic Statistics. Cambridge University Press.
van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes: with Applications to Statistics (Springer Series in Statistics). Springer.
van der Vaart, A.W., van Zanten, J.H., 2008. Rates of contraction of posterior distributions based on Gaussian process priors. Ann. Statist. 36, 1435–1463.
van der Vaart, A.W., van Zanten, J.H., 2009. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. Ann. Statist. 37, 2655–2675.
Villani, M., Kohn, R., Giordani, P., 2009. Regression density estimation using smooth adaptive Gaussian mixtures. J. Econometrics 153, 155–173.
Walker, S.G., 2004. Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist. 32, 2028–2043.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.
Wood, S., Jiang, W., Tanner, M., 2002. Bayesian mixture of splines for spatially adaptive nonparametric regression. Biometrika 89, 513–528.
Yau, P., Kohn, R., 2003. Estimation and variable selection in nonparametric heteroscedastic regression. Stat. Comput. 13, 191–208.
Zi-Zong, Y., 2009. Schur complements and determinant inequalities. J. Math. Inequal. 3, 161–167.