BAYESIAN REGRESSION WITH NONPARAMETRIC HETEROSKEDASTICITY

Andriy Norets

This paper presents a large sample justification for a semiparametric Bayesian approach to inference in a linear regression 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. It is shown that even when the data generating distribution of the error term is not normal the posterior distribution of the linear coefficients converges to a normal distribution with the mean equal to the asymptotically efficient estimator and the variance given by the semiparametric efficiency bound. This implies that the estimation procedure is robust and conservative from the Bayesian standpoint and at the same time it can be used as an implementation of semiparametrically efficient frequentist inference.

Keywords: Bayesian linear regression, heteroskedasticity, misspecification, posterior consistency, semiparametric Bernstein - von Mises theorem, semiparametric efficiency.

1. INTRODUCTION

This paper shows that a normal linear regression model with nonparametrically modeled heteroskedasticity is an attractive alternative to methods currently employed in the Bayesian econometric literature such as modeling the distribution of the error term by mixtures. Thus, it is argued that this model should be a more prominent part of the Bayesian toolbox for regression analysis.

Many different approaches to inference in a regression model have been proposed in the Bayesian framework. In the standard textbook linear regression model, the normality of the error terms is assumed. In the recent literature, the normality assumption is often relaxed by using mixtures of normal or Student t distributions for modeling the distribution of the errors. As pointed out by Mueller (2009), if the shape of the error distribution depends

[1]Current version: October 21, 2011.

[2]I am grateful to Bo Honore, Ulrich Mueller, Justinas Pelenis, Chris Sims, and Mark Watson for helpful discussions.

on covariates then the posterior may not concentrate around the data generating values of the linear coefficients. In the context of a linear regression model, Mueller (2009) suggests using the ordinary least squares estimator with the heteroskedasticity robust covariance matrix for Bayesian inference on the linear coefficients in possibly misspecified models. If the expectation of the response variable conditional on covariates is not linear in covariates and the linear regression provides a linear approximation to this conditional expectation, Mueller (2009)'s suggestion seems sound (it also has a pure Bayesian justification based on flexible multinomial-Dirichlet model or Bayesian bootstrap, see Lancaster (2003)). However, in the situations when the linearity of the conditional expectation is a reasonable assumption, the following approach seems to be more appropriate.

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, a transformation of splines or polynomials with a prior on parameters can be used as a prior for the variance. The normality of the error term guarantees that the Kullback-Leibler distance between the model and the data generating process (DGP), which does not necessarily satisfies the normality assumption, is minimized at the data generating values of the linear coefficients and the variance of the error term. Thus, one can expect the posterior consistency for these two parameters, which is proved in the paper.

Furthermore, the paper proves a Bernstein-von Mises type result for the linear coefficients: the posterior distribution of the linear coefficients converges to a normal distribution with the mean equal to the asymptotically efficient estimator and the variance given by the semiparametric efficiency bound. In the semiparametric efficiency literature, see, for example, Newey (1990) and Bickel et al. (1998), the model would be called a least favorable sub-model. The result suggests that the Bayesian inference about the linear coefficients based on this model is conservative in the following sense. Suppose we know the correct specification for the distribution of the error term and use it to estimate the linear coefficients. Then, the posterior variance in the correctly specified model cannot exceed the posterior variance in the least favorable normal model with the flexibly modeled error variance. Of course, one could go further and model the whole distribution of the error term flexibly in covariates with

the zero conditional mean restriction, see, for example, Pelenis (2010). It is also possible to model non-parametrically the distribution of the response conditional on covariates without imposing the linearity restriction, 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), De Iorio et al. (2004), Griffin and Steel (2006), Dunson and Park (2008), Chung and Dunson (2009), and Norets and Pelenis (2011) for models based on dependent Dirichlet processes. However, these more flexible models are harder to estimate and they require more data for reliable estimation results. In contrast, the model considered in the paper is parsimonious and at the same time it has attractive theoretical properties: consistent estimation of the error variance and linear coefficients and the conservativeness of the posterior distribution for the linear coefficients under misspecification. Thus, it can be thought of as a useful intermediate step between fully flexible models and simple models that could be inconsistent and misleading.

Bayesian Markov chain Monte Carlo (MCMC) estimation procedures for the normal regression with flexibly modeled variance have been developed in the literature, see, for example, Yau and Kohn (2003), who used transformed splines, or Goldberg et al. (1998), who used transformed Gaussian process prior for modeling the 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 estimators that achieve semiparametric efficiency such as Carroll (1982) and Robinson (1987). At the same time, the results of the paper provide a Bayesian interpretation to these classical semiparametrically efficient estimators.

The rest of the paper is organized as follows. Sections 2.1 and 2.2 describe the data generating process and the model. The Bernstein-von Mises theorem is presented in Section 2.3. Posterior consistency is considered in Section 2.4.

2. THEORETICAL RESULTS

2.1. *Data generating process and frequentist estimators*

The data are assumed to include n observations on a response variable and covariates $(Y^n, X^n) = (y_1, x_1, \ldots, y_n, x_n)$, where $y_i \in \mathcal{Y} \subset R$ and $x_i \in \mathcal{X} \subset R^k$, $i \in \{1, \ldots, n\}$. The observations are independently identically distributed (iid), $(y_i, x_i) \sim F_0$. The distribution of the infinite sequence of observations, (Y^∞, X^∞), is denoted by F_0^∞. The data generating process satisfies $E(y_i|x_i) = x_i'\beta_0$. Let $\epsilon_i = y_i - x_i'\beta_0$. Then, $E(\epsilon_i|x_i) = 0$. Assume $\sigma_0^2(x_i) = E(\epsilon_i^2|x_i)$ is well defined for any $x_i \in \mathcal{X}$. The joint DGP distribution F_0 is assumed to have a conditional density $f_0(y_i|x_i)$ with respect to the Lebesgue measure.

Chamberlain (1987) showed that the semiparametric efficiency bound for estimation of β_0 is given by $(E(x_i x_i' \sigma_0(x_i)^{-2}))^{-1}$. This is the asymptotic variance of the generalized least squares estimator under known σ_0,

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

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. A Bayesian analog of these results is derived below.

2.2. *Model and pseudo true parameter values*

The model postulates that $y_i|x_i \sim N(x_i'\beta, \sigma^2(x_i))$. The prior for $(\beta, \sigma(\cdot))$, Π, is a product of a normal prior for β, $N(\underline{\beta}, \underline{H}^{-1})$, and a distribution on a space of functions

$$\mathcal{S} \subset \{\sigma : \mathcal{X} \to [\underline{\sigma}, \overline{\sigma}]\}.$$

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

$$p(Y^n|X^n, \beta, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma(x_i)} \exp\left(-\frac{(y - x_i'\beta)^2}{2\sigma^2(x_i)} \right).$$

The posterior is given by

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

LEMMA 1 *Assume* $|\int \log f_0(y|x) dF_0(y, x)| < \infty$. *Then model parameter values* $\beta = \beta_0$ *and* $\sigma = \sigma_0$ *minimize the Kullback-Leibler (KL) distance between the DGP and the model.*

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) and White (1982) for classical results and Geweke (2005) for a textbook treatment of the Bayesian results). Posterior consistency in misspecified infinite dimensional models is discussed in Section 2.4 below.

2.3. *Bernstein-von Mises theorem*

The standard Bernstein-von Mises theorem shows that in well behaved parametric models the posterior distribution centered at the maximum likelihood estimator and scaled by $\sqrt{n}$ converges to a normal distribution with zero mean and a variance equal to the inverse of the Fisher information, see van der Vaart (1998) for a textbook treatment under weak regularity conditions. Thus, the theorem implies asymptotic equivalence between confidence and credible sets. 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 semiparametric Bernstein-von Mises theorem is an active area of current research, see, for example, Rivoirard and Rousseau (2009), Bickel and Kleijn (2010), and Castillo (2011). Misspecified semiparametric models are not covered by the existing results. The following theorem, which is the main result of the paper, is proved in Appendix A.

THEOREM 1 *Let* $d_2(\sigma_1^{-2}, \sigma_2^{-2}) = (\int [\sigma_1^{-2}(x) - \sigma_2^{-2}(x)]^2 dF_0(x))^{0.5}$. *Assume that*

1. *The marginal posterior of* σ *is consistent for the pseudo true value* σ_0*, i.e., for any* $\epsilon > 0$, $\Pi\left(d_2(\sigma_1^{-2}, \sigma_2^{-2}) > \epsilon \middle| Y^n, X^n\right) \to 0$ *in* F_0^∞ *probability.*
2. *For any* $x \in \mathcal{X}$, $\sigma_0(x) \in [\underline{\sigma}, \overline{\sigma}]$.
3. $0 < \underline{\sigma} < \overline{\sigma} < \infty$.
4. *For* $j = 1, \ldots, k$, $n^{-0.5} \sum x_{ij} \epsilon_i \sigma^{-2}(x_i)$ *converges weakly to a tight limit in the space of real bounded functions on* S *with the sup norm, where* x_{ij} *is coordinate* j *of* x_i.
5. $x_i x_i \sigma^{-2}(x_i)$, $\sigma \in S$, *is an* F_0*-Glivenko-Cantelli class of functions.*
6. *Vector* $x_i \epsilon_i$ *has finite second moments.*
7. x_i *has finite fourth moments.*
8. $E[x_i x_i' \sigma_0(x_i)^{-2}]$ *exists and it is invertible.*

Then, the total variation distance

$$d_{TV}\left(\Pi[n^{0.5}(\beta - \hat{\beta}_{GLS})|Y^n, X^n], N\left(0, \left(E[x_i x_i' \sigma_0(x_i)^{-2}]\right)^{-1}\right)\right) \to 0 \tag{1}$$

in F_0^∞ *probability.*

Lemma 2.3.11 in van der Vaart and Wellner (1996) implies the following sufficient conditions for the weak convergence assumed in condition 4 of the theorem: (S, d_2) is totally bounded and $n^{-0.5} \sum x_{ij} \epsilon_i \sigma^{-2}(x_i)$ is stochastically equicontinuous in (S, d_2).[1] Total boundedness of (S, d_2) is also essential for the Glivenko-Cantelli class assumption. Andrews (1986), pages 2175 and 2171, provides a set of sufficient conditions for stochastic equicontinuity: existence of $2 + \delta$ moments for $x_i \epsilon_i$, where $\delta > 0$, and existence of uniformly bounded and uniformly Lipschitz continuous partial derivatives of order at least $k/2$ for functions in S.

Posterior consistency for σ is considered in the next section under the assumption of bounded prior support of β. Thus, the following corollary proved in Appendix B is useful.

COROLLARY 1 *Theorem 1 remains true if the normal prior for* β *is truncated to a set* $[-B, B]^k$*, with a sufficiently large* $B > 0$.

[1] ρ_Z metric in Lemma 2.3.11 of van der Vaart and Wellner (1996) is dominated by d_2 under the assumptions of Theorem 1 and existence of $2 + \delta$ moments for $x_i \epsilon_i$, where $\delta > 0$.

2.4. *Posterior consistency*

Posterior consistency in correctly specified semi- and non-parametric models is well understood, see Ghosh and Ramamoorthi (2003) for a textbook treatment. Available extensions of the posterior consistency arguments to misspecified non-parametric models are much more involved than the corresponding extensions in the parametric case and the sufficient conditions seem to be rather strong. For example, the sufficient conditions in Kleijn and van der Vaart (2006) applied to the normal linear heteroskedastic model similarly to their Section 4 seem to rule out DGPs with normally distributed ϵ_i. Thus, the proof of the posterior consistency result presented below is model specific.

Since $\sigma \in S$ are uniformly bounded above and away from zero, the $L_2(F_0)$ distance, $d_2(\sigma_1^{-2}, \sigma_2^{-2})$ used in the previous section is equivalent to $d_2(\sigma_1^2, \sigma_2^2)$ and the latter is used to define a distance on the whole parameter space

$$\rho_2\big((\beta_1, \sigma_1^2), (\beta_2, \sigma_2^2)\big)^2 = E[(\sigma(x_i)^2 - \sigma_0(x_i)^2)^2] + ||\beta - \beta_0||_2^2.$$

THEOREM 2 *Assume that*

1. *Prior puts positive probability on any ρ_2 neighborhood of (β_0, σ_0^2), i.e., for any $\epsilon > 0$,* $\Pi\big(\sigma, \beta : \ \rho_2((\beta, \sigma^2), (\beta_0, \sigma_0^2)) < \epsilon\big) > 0$.
2. *Prior for β has bounded support, $[-B, B]^k$.*
3. *$\{\epsilon_i^2 \sigma^{-2}(x_i), \sigma \in S\}$, $\{\log(\sigma^2(x_i)), \sigma \in S\}$, and $\{x_i \epsilon_i \sigma^{-2}(x_i), \sigma \in S\}$ are F_0-Glivenko-Cantelli classes.*
4. *$E(x_i x_i')$ is invertible.*

Then, the posterior is consistent in ρ_2: for $U = [\rho_2\big((\beta, \sigma^2), (\beta_0, \sigma_0^2)\big) \leq \epsilon]$

$$\Pi\left[U^c \middle| Y^n, X^n\right] \to 0$$

in F_0^∞ probability.

Priors for σ that satisfy the assumptions in Theorems 1-2 can be based on transformations of splines or polynomials with additional bounds on derivatives discussed after Theorem 1. Theorems 1-2 can also be formulated for a sample size dependent prior, Π_n. In this case, the

assumptions of the theorems such as Glivenko-Cantelli properties have to hold on a set S_n. The prior probability, $\Pi_n(S_n^c)$, should converge to zero sufficiently fast so that the posterior probability $\Pi_n(S_n^c|Y^n, X^n)$ converges to zero. Such a generalization can be used to relax the assumptions of uniformly bounded derivatives and boundedness above and away from zero for σ.

3. APPENDIX A. PROOFS OF MAIN RESULTS

PROOF: Theorem 1.

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

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

Derivations of conditional posteriors in linear regression models can be found in any Bayesian textbook, see, for example, Geweke (2005). The (marginal) posterior of β can be expressed as

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

After normalization $z = n^{0.5}(\beta - \hat{\beta}_{GLS})$, the conditional posterior is still normal

$$\Pi(z|\sigma, Y^n, X^n) = \phi\left(z, n^{0.5}(\overline{\beta} - \hat{\beta}_{GLS}), \left(\overline{H}/n\right)^{-1}\right),$$

where $\phi(\cdot, \cdot, \cdot)$ denotes the density of the normal distribution.

The total variation distance of interest can be expressed as follows

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

To bound the total variation distance between the two normal distributions inside the last integral one can use the following two facts. First, the total variation distance is bounded by 2 times the square root of the KL distance, see for example, Proposition 1.2.2 in Ghosh and

Ramamoorthi (2003). Second, the KL distance between two normal distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ is equal to

$$\frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} + \text{tr}(\Sigma_2^{-1}\Sigma_1 - I) + (\mu_1 - \mu_2)'\Sigma_2^{-1}(\mu_1 - \mu_2)\right)$$
$$\leq \frac{\left||\Sigma_2^{-1}| - |\Sigma_1^{-1}|\right|}{\min(|\Sigma_2^{-1}|, |\Sigma_1^{-1}|)} + k \cdot ||\Sigma_2^{-1} - \Sigma_1^{-1}||_\infty \cdot ||\Sigma_1||_\infty + ||\mu_1 - \mu_2||_2^2 \cdot ||\Sigma_2^{-1}||_2,$$

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 R^k, $||\mu||_2^2 = \sum_{i=1}^k \mu_i^2$. Thus,

$$d_{TV}\left(\Pi[n^{0.5}(\beta - \hat{\beta}_{GLS})|Y^n, X^n], N\left(0, \left(E[x_i x_i' \sigma_0(x_i)^{-2}]\right)^{-1}\right)\right)$$
$$\leq 2\int \sqrt{A_n + B_n + C_n} d\Pi(\sigma|Y^n, X^n),$$

where

$$A_n = \frac{\left||H/n| - |E[x_i x_i' \sigma_0(x_i)^{-2}]|\right|}{\min(|H/n|, |E[x_i x_i' \sigma_0(x_i)^{-2}]|)}$$
$$B_n = k \cdot ||H/n - E[x_i x_i' \sigma_0(x_i)^{-2}]||_\infty \cdot ||\left(E[x_i x_i' \sigma_0(x_i)^{-2}]\right)^{-1}||_\infty$$
$$C_n = ||H/n||_2 \cdot ||\left(\frac{1}{n}\sum_i \frac{x_i x_i'}{\sigma_0(x_i)^2}\right)^{-1} \frac{1}{\sqrt{n}}\sum_i \frac{x_i y_i}{\sigma_0(x_i)^2}$$
$$- (H/n)^{-1}\left(\frac{H\beta}{\sqrt{n}} + \frac{1}{\sqrt{n}}\sum_i \frac{x_i y_i}{\sigma(x_i)^2}\right)||_2^2.$$

By Lemmas 2-4 and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any nonnegative a and b, it suffices to show that $\int d_2(\sigma_0^{-2}, \sigma^{-2}) d\Pi(\sigma|Y^n, X^n)$ and $\int ||n^{-0.5}\sum_i x_i \epsilon_i(\sigma_0(x_i)^{-2} - \sigma(x_i)^{-2}||_2 d\Pi(\sigma|Y^n, X^n)$ converge to zero in outer probability F_0^∞. For a definition of outer probability see, for example, van der Vaart and Wellner (1996). It is usually introduced to avoid possible measurability issues for supremums over large classes of functions. Convergence in probability in the proof of theorems and auxiliary results below should be understood as convergence in outer probability whenever necessary. Because convergence in outer probability is established for components of an upper bound on the total variation distance in (1), expression in (1) will converge in probability as long as it is measurable, which is clearly the case.

Since $d_2(\sigma_0^{-2},\sigma^{-2}) \leq \underline{\sigma}^{-2}$,

$$F_0^\infty \left[\int d_2(\sigma_0^{-2},\sigma^{-2}) d\Pi(\sigma|Y^n,X^n) > \epsilon \right] \leq F_0^\infty \left[\underline{\sigma}^{-2} \Pi(d_2(\sigma_0^{-2},\sigma^{-2}) > \epsilon/2 | Y^n,X^n) + \epsilon/2 > \epsilon \right]$$
$$= F_0^\infty \left[\Pi(d_2(\sigma_0^{-2},\sigma^{-2}) > \epsilon/2 | Y^n,X^n) > \epsilon/(2\underline{\sigma}^{-2}) \right] \to 0, \ \forall \epsilon > 0.$$

By Lemma 5 there exists a positive sequence $\delta_n \to 0$ such that $\Pi(d_2(\sigma_0^{-2},\sigma^{-2}) > \delta_n | Y^n,X^n)$ converges to zero in probability. Then,

$$\int ||n^{-0.5} \sum_i x_i \epsilon_i (\sigma_0(x_i)^{-2} - \sigma(x_i)^{-2}||_2 d\Pi(\sigma|Y^n,X^n) \tag{3}$$
$$\leq \sup_{\sigma \in S} ||n^{-0.5} \sum_i x_i \epsilon_i (\sigma_0^{-2}(x_i) - \sigma^{-2}(x_i)||_2 \cdot \Pi(d_2(\sigma_0^{-2},\sigma^{-2}) > \delta_n | Y^n,X^n)$$
$$+ \sup_{\sigma: d_2(\sigma_0^{-2},\sigma^{-2}) \leq \delta_n} ||n^{-0.5} \sum_i x_i \epsilon_i (\sigma_0^{-2}(x_i) - \sigma^{-2}(x_i)||_2.$$

By Lemma 2.3.9 in van der Vaart and Wellner (1996), the assumed finiteness of the second moments of $x_i\epsilon$, and the assumed weak convergence of $n^{-0.5}\sum_i x_i \epsilon_i \sigma^{-2}(x_i)$,

$$\sup_{\sigma \in S} ||n^{-0.5} \sum_i x_i \epsilon_i (\sigma_0^{-2}(x_i) - \sigma^{-2}(x_i)||_2$$

is bounded in probability. Thus, the first part of the bound in (3) converges to zero in probability. By Lemma 2.3.11 in van der Vaart and Wellner (1996) and the assumed weak convergence, the second part of the bound in (3) converges to zero in outer probability.

Q.E.D.

PROOF: Theorem 2.

Let $p_{\beta,\sigma,i}$ denote a normal density with mean $x_i'\beta$ and variance $\sigma^2(x_i)$. The posterior can be expressed as

$$\Pi\left(U^c \middle| Y^n,X^n\right) = \frac{\int_{U^c} \prod_{i=1}^n p_{\beta,\sigma,i}/p_{\beta_0,\sigma_0,i} d\Pi(\beta,\sigma)}{\int \prod_{i=1}^n p_{\beta,\sigma,i}/p_{\beta_0,\sigma_0,i} d\Pi(\beta,\sigma)}$$
$$= \frac{\exp(n\delta) \int_{U^c} \exp\{\sum_{i=1}^n \log(p_{\beta,\sigma,i}/p_{\beta_0,\sigma_0,i})\} d\Pi(\beta,\sigma)}{\exp(n\delta) \int \exp\{\sum_{i=1}^n \log(p_{\beta,\sigma,i}/p_{\beta_0,\sigma_0,i})\} d\Pi(\beta,\sigma)}. \tag{4}$$

Thus, it suffices to show that the numerator converges to zero in probability F_0^∞ for some $\delta > 0$ and the denominator converges to infinity a.s. F_0^∞ for any $\delta > 0$. Consider the

numerator first.

$$\frac{1}{n}\sum_i \log\frac{p_{\beta_0,\sigma_0,i}}{p_{\beta,\sigma,i}} = \frac{1}{2}\frac{1}{n}\sum_i \Bigg(\log\frac{\sigma^2(x_i)}{\sigma_0^2(x_i)} - \frac{\epsilon_i^2}{\sigma_0^2(x_i)} + \frac{\epsilon_i^2}{\sigma^2(x_i)} + 2\frac{\epsilon_i x_i'}{\sigma^2(x_i)}(\beta_0-\beta)$$
$$+ (\beta-\beta_0)'\left(\frac{x_i x_i'}{\sigma^2(x_i)}\right)(\beta-\beta_0)\Bigg)$$
$$= \frac{1}{2}\Bigg[\frac{1}{n}\sum_i \log\sigma^2(x_i) - E(\log\sigma^2(x_i)) + E(\log\sigma_0^2(x_i)) - \frac{1}{n}\sum_i \log\sigma_0^2(x_i)$$
$$+ E\left(\frac{\epsilon_i^2}{\sigma_0^2(x_i)}\right) - \frac{1}{n}\sum_i \frac{\epsilon_i^2}{\sigma_0^2(x_i)} + \frac{1}{n}\sum_i \frac{\epsilon_i^2}{\sigma^2(x_i)} - E\left(\frac{\epsilon_i^2}{\sigma^2(x_i)}\right) + \frac{2}{n}\sum \frac{\epsilon_i x_i'}{\sigma^2(x_i)}(\beta_0-\beta)$$
$$+ E\left(\log\frac{\sigma^2(x_i)}{\sigma_0^2(x_i)} - \frac{\epsilon_i^2}{\sigma_0^2(x_i)} + \frac{\epsilon_i^2}{\sigma^2(x_i)}\right) + (\beta-\beta_0)'\left(\frac{1}{n}\sum \frac{x_i x_i'}{\sigma^2(x_i)}\right)(\beta-\beta_0)\Bigg].$$

Lines 3 and 4 of the preceding display converge to zero uniformly over $\sigma \in S$ by the (Glivenko-Cantelli class) assumptions of the theorem and boundedness of $||\beta_0 - \beta||_2$. Thus, they can be bounded below by some Q_n, which does not depend on σ and converges to zero in outer probability. The last line can be bounded below by

$$E\left(\log(\sigma^2(x_i)/\sigma_0^2(x_i)) - 1 + \sigma_0^2(x_i)/\sigma^2(x_i)\right) + \lambda_n||\beta-\beta_0||_2^2$$
$$\geq \min(\lambda_n, C)\cdot\rho_2^2\big((\beta,\sigma^2),(\beta_0,\sigma_0^2)\big),$$

where λ_n is the smallest eigenvalue of $\sum x_i x_i'/(n\overline{\sigma}^2)$, which converges in probability to a positive limit λ, and constant C is defined in (15) in the proof of Lemma 7. Since $\rho_2((\beta,\sigma^2),(\beta_0,\sigma_0^2)) > \epsilon$ over U^c, the numerator in (4) can be bounded above by $\exp\{n(\delta - 0.5Q_n - 0.5\min(\lambda_n, C)\cdot\epsilon^2)\}$, which converges to zero in outer probability for $\delta < 0.25\min(\lambda, C)\cdot\epsilon^2$.

Let us consider the denominator. The assumption of positive prior probability of ρ_2 neighborhoods of (β_0, σ_0) (theorem's condition 1) and Lemma 7 imply that for any $\delta > 0$,

$$\Pi(\beta,\sigma : E(\log(p_{\beta_0,\sigma_0,i}/p_{\beta,\sigma,i})) < \delta) > 0.$$

Then an argument based on Fubini's theorem and Fatou's lemma that shows that the denominator converges to infinity a.s. in the proof of Schwartz posterior consistency theorem (see, Ghosh and Ramamoorthi (2003), pp. 129-130) applies without any changes. [2] *Q.E.D.*

[2] Unfortunately, an analogous argument based on the lim sup version of Fatou's lemma does not apply

4. APPENDIX B. PROOFS OF AUXILIARY RESULTS

PROOF: Lemma 1.

The marginal distribution of x is canceled out inside log in the KL distance

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

Since $E[(y-x'\beta)^2|x] = \sigma_0(x)^2 + [x'(\beta-\beta_0)]^2$, $\beta=\beta_0$ is the minimizer of the KL distance for any function σ. Then, it follows immediately from the first order conditions that $\sigma=\sigma_0$ minimizes the KL distance. *Q.E.D.*

LEMMA 2 *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 $B_n^1 \overset{F_0^\infty}{\to} 0$ and $B_n^2 \overset{F_0^\infty}{\to} B^2$, B^2 is a constant, and (B_n^1, B_n^2) do not depend on σ.*

PROOF:

$$||H/n - E[x_i x_i' \sigma_0(x_i)^{-2}]||_\infty \leq ||\underline{H}/n||_\infty$$
$$+ \sup_{\sigma\in S} ||\frac{1}{n}\sum_i \frac{x_i x_i'}{\sigma^2(x_i)} - E\left(\frac{x_i x_i'}{\sigma^2(x_i)}\right)||_\infty + ||E\left(x_i x_i'\left(\frac{1}{\sigma^2(x_i)} - \frac{1}{\sigma_0^2(x_i)}\right)\right)||_\infty.$$

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 $d_2(\sigma^{-2},\sigma_0^{-2})$. *Q.E.D.*

LEMMA 3 *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 $A_n^1 \overset{F_0^\infty}{\to} 0$ and $A_n^2 \overset{F_0^\infty}{\to} A^2$, 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 $k\times k$ matrices A and B, $\big||A| - |B|\big| \leq k!\cdot k \max(||A||_\infty, ||B||_\infty)^{k-1}\cdot||A-B||_\infty$. Thus, the

to the numerator as the lim sup version requires an integrable upper bound. Thus, assumptions similar to theorem's condition 3 are needed to handle the numerator.

numerator of A_n is bounded by a multiple of the bound on B_n derived in Lemma 2 times $\max(||H/n||_\infty, ||E[x_i x_i' \sigma_0(x_i)^{-2}]||_\infty)^{k-1}$. 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 $|\sum_i x_i x_i'/n| \overset{F_0^\infty}{\to} |E[x_i x_i']| > 0$, the claim of the lemma follows. *Q.E.D.*

LEMMA 4 *The following inequality holds for C_n defined in the proof of Theorem 1*

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

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

PROOF: Plugging $y = x_i'\beta_0 + \epsilon_i$ into the definition of C_n results in

$$\begin{aligned} \sqrt{C_n/||H/n||_2} &= || \, (H/n)^{-1} \, \underline{H}(\beta_0 - \underline{\beta})/\sqrt{n} \\ &+ \left(\frac{1}{n} \sum_i \frac{x_i x_i'}{\sigma_0(x_i)^2} \right)^{-1} \frac{1}{\sqrt{n}} \sum_i \frac{x_i \epsilon_i}{\sigma_0(x_i)^2} - (H/n)^{-1} \frac{1}{\sqrt{n}} \sum_i \frac{x_i \epsilon_i}{\sigma(x_i)^2} ||_2. \end{aligned} \tag{5}$$

The first expression on the right hand side of (5) 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 (8)). The norm of the second expression can be bounded by[3]

$$\begin{aligned} &|| \, (H/n)^{-1} \, ||_2 \cdot || \frac{1}{\sqrt{n}} \sum_i x_i \epsilon_i \left(\frac{1}{\sigma_0(x_i)^2} - \frac{1}{\sigma(x_i)^2} \right) ||_2 \\ &+ || \left(\frac{1}{n} \sum_i \frac{x_i x_i'}{\sigma_0(x_i)^2} \right)^{-1} - (H/n)^{-1} \, ||_2 \cdot || \frac{1}{\sqrt{n}} \sum_i \frac{x_i \epsilon_i}{\sigma_0(x_i)^2} \, ||_2. \end{aligned} \tag{6}$$

The norm of the difference in the inverses in the second line of (6) is bounded by[4]

$$|| \left(\frac{1}{n} \sum_i \frac{x_i x_i'}{\sigma_0(x_i)^2} \right)^{-1} ||_2 \cdot || \, (H/n)^{-1} \, ||_2 \cdot || (\sum_i x_i x_i' (\sigma_0(x_i)^{-2} - \sigma(x_i)^{-2}) - \underline{H})/n ||_2. \tag{7}$$

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

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

Next, we separately consider the three parts of the product in (7). The first part converges to $||\left(E(x_i x_i' \sigma_0(x_i)^{-2})\right)^{-1}||_2$ in probability. The second part,

$$
\begin{aligned}
(8) \qquad & ||(H/n)^{-1}||_2 = \sup_x \frac{||(H/n)^{-1}x||_2}{||x||_2} = \sup_x \frac{||(H/n)^{-1}(H/n)y||_2}{||(H/n)y||_2} \\
& = \left(\inf_y \frac{||(H/n)y||_2}{||y||_2}\right)^{-1} = \left(\inf_y \frac{||y||_2 \cdot ||(H/n)y||_2}{||y||_2^2}\right)^{-1} \\
& \leq \left(\inf_y \frac{|y'(H/n)y|}{||y||_2^2}\right)^{-1} \leq \left(\inf_y \frac{|y'((\underline{H}\overline{\sigma} + \sum_i x_i x_i')/n)y|/\overline{\sigma}}{||y||_2^2}\right)^{-1} \\
& = \frac{\overline{\sigma}}{\lambda_{\min}((\underline{H}\overline{\sigma} + \sum_i x_i x_i')/n)} \overset{F_0^\infty}{\to} \frac{\overline{\sigma}}{\lambda_{\min}(E(x_i x_i'))},
\end{aligned}
$$

where $\lambda_{\min}(\cdot)$ 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[5].

The third part of the product in (7) is bounded above by

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

which can be bounded as in Lemma 2 ($||A||_2 \leq \dim(A)||A||_\infty$). The bounds derived above and the Slutsky theorem imply the claim of the lemma.

Q.E.D.

LEMMA 5 *If for any $\epsilon > 0$ $\Pi\left(||\sigma - \sigma_0|| > \epsilon \middle| Y^n, X^n\right) \to 0$ in F_0^∞ probability then for any positive constants a, b, c, and d there exists a sequence $\{\epsilon_n\} \to 0$ such that*

$$
F_0^\infty\left(\Pi\left[||\sigma - \sigma_0|| > a\epsilon_n \middle| Y^n, X^n\right] > b/\sqrt{c + d/\epsilon_n}\right) \to 0.
$$

PROOF: By the definition of convergence in probability, for any fixed m there exists N_m such that for any $n \geq N_m$

$$
F_0^\infty\left(\Pi\left(||\sigma - \sigma_0|| > a/m \middle| Y^n, X^n\right) > b/\sqrt{c + dm}\right) < 1/m.
$$

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

The sequence of N_m can be chosen to be increasing. Then, for $n \in [N_m, N_{m+1})$ set $\epsilon_n = 1/m$. *Q.E.D.*

LEMMA 6 *For two distributions P_1 and P_2 with densities p_1 and p_2 with respect to a measure μ, the total variation distance between P_2 truncated to a set E and P_1 can be bounded as follows*

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

PROOF:

$$\begin{aligned}
&\int |p_1 - 1_E p_2 / P_2(E)| = \int_E |p_1 P_2(E) - p_2| / P_2(E) + P_1(E^c) \\
&\leq \int_E |p_1(P_2(E) - 1) + p_1 - p_2| / P_2(E) + P_1(E^c) \\
&\leq P_1(E^c) + (1 - P_2(E))/P_2(E) + \int |p_1 - p_2| / P_2(E).
\end{aligned}$$

Q.E.D.

PROOF: Corollary 1.

The proof is a slight modification of the proof of Theorem 1. With the truncated prior, the conditional posterior of $z = \sqrt{n}(\beta - \hat{\beta}_{GLS})$, $\Pi(z|\sigma, Y^n, X^n)$, is $N(\sqrt{n}(\overline{\beta} - \hat{\beta}_{GLS}), (\overline{H}/n)^{-1})$ truncated to $\sqrt{n}([-B, B]^k - \hat{\beta}_{GLS})$. Thus, with the truncated prior, the total variation distance in the last line of (2) is a distance between a truncated normal distribution and a normal distribution. By Lemma 6 and the proof of Theorem 1, it suffices to show that the probability of set $\sqrt{n}([-B, B]^k - \hat{\beta}_{GLS})$ under $N(\sqrt{n}(\overline{\beta} - \hat{\beta}_{GLS}), (\overline{H}/n)^{-1})$ and $N(0, (E[x_i x_i' \sigma_0(x_i)^{-2}])^{-1})$ is bounded below by a bound that does not depend on σ and converges to 1 in F_0^∞ probability.

$$\begin{aligned}
&1 - \int_{\sqrt{n}([-B,B]^k - \hat{\beta}_{GLS})} \phi(z, (\sqrt{n}(\overline{\beta} - \hat{\beta}_{GLS}), (\overline{H}/n)^{-1}) dz \\
&= 1 - \int_{\sqrt{n}([-B,B]^k - \overline{\beta})} \phi(z, 0, (\overline{H}/n)^{-1}) dz \\
&\leq \sum_{i=1}^k \int_{z_i \notin \sqrt{n}([-B,B] - \overline{\beta}_i)} \phi(z, 0, (\overline{H}/n)^{-1}) dz.
\end{aligned}$$

Next, note that

$$z'(\overline{H}/n)z \geq z'[(\underline{H} + \sum_i x_i x_i'/\overline{\sigma}^2)/n]z \geq z'z\lambda_m^n, \tag{9}$$

where λ_m^n is the smallest eigenvalue of $(\underline{H} + \sum_i x_i x_i'/\overline{\sigma}^2)/n)$. Also, as in the proof of Lemma 3,

$$|\overline{H}/n| \leq |(\underline{H} + \sum_i x_i x_i'/\underline{\sigma}^2)/n|. \tag{10}$$

Using the bound on $||(\overline{H}/n)^{-1})||_2$ from Lemma 4, we get

$$\begin{aligned} |\overline{\beta}_i| \leq ||\overline{\beta}||_2 \leq ||(\overline{H}/n)^{-1})||_2 \cdot ||(\underline{H}\underline{\beta} + \sum_i |x_i y_i|/\underline{\sigma}^2)/n||_2 \qquad (11) \\ \leq \frac{\overline{\sigma}||(\underline{H}\underline{\beta} + \sum_i |x_i y_i|/\underline{\sigma}^2)/n||_2}{\lambda_{\min}((\underline{H}\overline{\sigma} + \sum_i x_i x_i')/n)} \equiv F_n \overset{F_0^\infty}{\to} F \equiv \frac{\overline{\sigma}||E|x_i y_i|/\underline{\sigma}^2||_2}{\lambda_{\min}(E(x_i x_i'))}. \end{aligned}$$

The assumption of the corollary that B is sufficiently large means that $B > F$.

From (9) - (11),

$$\begin{aligned} &\int_{z_i \notin \sqrt{n}([-B,B]-\overline{\beta}_i)} \phi(z, 0, (\overline{H}/n)^{-1})dz \leq 2\int_{z_i \geq \sqrt{n}(B-|\overline{\beta}_i|)} \phi(z, 0, (\overline{H}/n)^{-1})dz \\ &\leq 2|(\underline{H} + \sum_i x_i x_i'/\underline{\sigma}^2)/n|^{0.5} \int_{z_i \geq \sqrt{n}(B-F_n)} \exp\{-0.5 z'z\lambda_m^n\}(2\pi)^{-k/2}dz \\ &\leq 2|(\underline{H} + \sum_i x_i x_i'/\underline{\sigma}^2)/n|^{0.5}(\lambda_m^n)^{-k/2} \int_{z_i \geq \sqrt{n}(B-F_n)\lambda_m^n} \exp\{-0.5 z_i^2\}(2\pi)^{-1/2}dz. \end{aligned}$$

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

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

where the convergence in probability follows from the convergence of F_n and λ_m^n. This completes the proof of convergence for the probability of $\sqrt{n}([-B, B]^k - \hat{\beta}_{GLS})$ under $N(\sqrt{n}(\overline{\beta} - \hat{\beta}_{GLS}), (\overline{H}/n)^{-1})$. The proof for $N(0, (E[x_i x_i'\sigma_0(x_i)^{-2}])^{-1})$ is similar. *Q.E.D.*

LEMMA 7 *For some positive constants C_0 and C_1*

$$E(\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma})) \geq C_0\rho_2^2\big((\beta, \sigma^2), (\beta_0, \sigma_0^2)\big), \tag{12}$$

$$E(\log(p_{\beta_0,\sigma_0}/p_{\beta,\sigma})) \leq C_1\rho_2^2\big((\beta, \sigma^2), (\beta_0, \sigma_0^2)\big). \tag{13}$$

PROOF: The law of iterated expectations implies

$$E(\log \frac{p_{\beta_0,\sigma_0}}{p_{\beta,\sigma}}) = \frac{1}{2} E\left(\log \frac{\sigma^2(x_i)}{\sigma_0^2(x_i)} + \frac{\sigma_0^2(x_i) - \sigma^2(x_i)}{\sigma^2(x_i)} + (\beta - \beta_0)' \left(\frac{x_i x_i'}{\sigma^2(x_i)} \right) (\beta - \beta_0) \right). \tag{14}$$

First, note that

$$\frac{\lambda_{\min}(E(x_i x_i'))}{\overline{\sigma}^2} ||\beta - \beta_0||_2^2 \leq (\beta - \beta_0)' E\left(\frac{x_i x_i'}{\sigma^2(x_i)} \right) (\beta - \beta_0) \leq \frac{\lambda_{\max}(E(x_i x_i'))}{\underline{\sigma}^2} ||\beta - \beta_0||_2^2,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues. 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 $z \in [\underline{\sigma}^2/\overline{\sigma}^2, \overline{\sigma}^2/\underline{\sigma}^2]$, $q(\overline{\sigma}^2/\underline{\sigma}^2) \leq q(z) \leq q(\underline{\sigma}^2/\overline{\sigma}^2)$. From this inequality,

$$\frac{E(\sigma_0^2 - \sigma^2)^2}{\overline{\sigma}^4} q(\overline{\sigma}^2/\underline{\sigma}^2) \leq E\left(\log \frac{\sigma^2}{\sigma_0^2} + \frac{\sigma_0^2 - \sigma^2}{\sigma^2} \right) \leq \frac{E(\sigma_0^2 - \sigma^2)^2}{\underline{\sigma}^4} q(\underline{\sigma}^2/\overline{\sigma}^2). \tag{15}$$

Thus, inequalities (12) and (13) are proved. *Q.E.D.*

REFERENCES

ANDREWS, D. W. (1986): "Empirical process methods in econometrics," in *Handbook of Econometrics*, ed. by R. F. Engle and D. McFadden, Elsevier, vol. 4 of *Handbook of Econometrics*, chap. 37, 2247–2294.

BICKEL, P. AND B. KLEIJN (2010): "The semiparametric Bernstein-Von Mises theorem," *ArXiv e-prints*.

BICKEL, P. J., C. A. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1998): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer.

CARROLL, R. J. (1982): "Adapting for Heteroscedasticity in Linear Models," *The Annals of Statistics*, 10, pp. 1224–1233.

CASTILLO, I. (2011): "A semi-parametric Bernstein-von Mises theorem for Gaussian process priors," To appear in Probability Theory and Related Fields.

CHAMBERLAIN, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

CHUNG, Y. AND D. B. DUNSON (2009): "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection," *Journal of the American Statistical Association*, 104, 1646–1660.

DE IORIO, M., P. MULLER, G. L. ROSNER, AND S. N. MACEACHERN (2004): "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215.

Dunson, D. B. and J.-H. Park (2008): "Kernel stick-breaking processes," *Biometrika*, 95, 307–323.

Geweke, J. (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience.

Geweke, J. and M. Keane (2007): "Smoothly mixing regressions," *Journal of Econometrics*, 138, 252–290.

Ghosh, J. and R. Ramamoorthi (2003): *Bayesian Nonparametrics*, Springer; 1 edition.

Goldberg, P. W., C. K. I. Williams, and C. M. Bishop (1998): "Regression with Input-dependent Noise: A Gaussian Process Treatment," in *In Advances in Neural Information Processing Systems 10*, MIT Press, 493–499.

Griffin, J. E. and M. F. J. Steel (2006): "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

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*, Berkeley: University of California Press, vol. 1, 221–233.

Kleijn, B. and A. van der Vaart (2006): "Misspecification in Infinite-Dimensional Bayesian Statistics," *The Annals of Statistics*, 34, 837–877.

Lancaster, T. (2003): "A Note on Bootstraps and Robustness," .

MacEachern, S. N. (1999): "Dependent Nonparametric Processes," *ASA Proceedings of the Section on Bayesian Statistical Science*.

Mueller, U. (2009): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix," .

Newey, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

Norets, A. (2010): "Approximation of conditional densities by smooth mixtures of regressions," *The Annals of statistics*, 38, 1733–1766.

Norets, A. and J. Pelenis (2011): "Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures," Unpublished manuscript, Princeton University.

Pelenis, J. (2010): "Bayesian Semiparametric Regression," Unpublished manuscript, Princeton University.

Peng, F., R. A. Jacobs, and M. A. Tanner (1996): "Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With an Application to Speech Recognition," *Journal of the American Statistical Association*, 91, 953–960.

Rivoirard, V. and J. Rousseau (2009): "Bernstein Von Mises Theorem for linear functionals of the density," *ArXiv e-prints*.

Robinson, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

Shen, X. (2002): "Asymptotic Normality of Semiparametric and Nonparametric Posterior Distributions," *Journal of the American Statistical Association*, 97, 222–235.

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

VAN DER VAART, A. AND J. WELLNER (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics (Springer Series in Statistics)*, Springer.

VILLANI, M., R. KOHN, AND P. GIORDANI (2009): "Regression density estimation using smooth adaptive Gaussian mixtures," *Journal of Econometrics*, 153, 155 – 173.

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, pp. 1–25.

WOOD, S., W. JIANG, AND M. TANNER (2002): "Bayesian mixture of splines for spatially adaptive nonparametric regression," *Biometrika*, 89, 513–528.

YAU, P. AND R. KOHN (2003): "Estimation and variable selection in nonparametric heteroscedastic regression," *Statistics and Computing*, 13, 191–208.

ZI-ZONG, Y. (2009): "Schur Complements and Determinant Inequalities," *Journal of Mathematical Inequalities*, 3, 161–167.