Submitted to

LOCALLY ROBUST EFFICIENT BAYESIAN INFERENCE

ULRICH MÜLLER
Economics Department, Princeton University

ANDRIY NORETS
Economics Department, Brown University

We propose a framework for making Bayesian parametric models robust to local misspecification. Suppose in a baseline parametric model, a parameter of interest has an interpretation in a more general semiparametric model and the baseline model is only locally misspecified. In general, Bayesian and maximum likelihood estimators will be asymptotically biased in these settings. We propose to augment the baseline likelihood by a multiplicative factor that involves scores for the baseline model, the efficient scores for the encompassing semiparametric model, and an auxiliary parameter that has the same dimension as the parameter of interest. We show that this augmented model results in a marginal posterior for the parameter of interest that is asymptotically normal with mean equal to the semiparametrically efficient estimator and variance equal to the semiparametric efficiency bound. The suggested augmentation thus robustifies the baseline parametric model to local misspecification, while preserving the appeal of Bayesian inference. We develop an MCMC algorithm for the estimation of the augmented model, and illustrate the approach in applications.

KEYWORDS: Bayesian methods, Semiparametric efficiency, Bernstein-von Mises theorem, Local misspecification, Robustness.

1. INTRODUCTION

Consider a researcher seeking to conduct Bayesian inference in a simple location model with independently identically distributed (i.i.d.) observations. The researcher is interested both in the population mean, and the quantiles of the distribution (say, for forecasting purposes). The data seems symmetric, but with tails that are heavier than those of a normal model. The researcher thus follows textbook advice and models the data as distributed Student's t, shifted by the location parameter.

By the parametric Bernstein-von Mises theorem, if the Student's t model is correct, the large sample posterior for the population mean is approximately normal with the same asymptotic variance as the maximum likelihood estimator (MLE). This variance is smaller than the variance of the sample mean. Yet, as is well known, the sample mean is the semiparametrically efficient estimator of the location parameter. By implication, there exist local deviations of the Student's t model that induce a local bias in the MLE, and thus the posterior distribution, that are of the same order as the posterior uncertainty about the population mean. These deviations are not detectable with probability close to unity, even in large samples. So the researcher has no way of knowing for sure that the Student's t model is misspecified, and the implications of the Student's t model for the data quantiles continue to be correct to first order.

Of course, if the researcher is confident in the correctness of the Student's t model, then these considerations are irrelevant. But if the Student's t model was merely chosen for convenience and analytical tractability, then they are potentially worrying: implicitly, the Student's t model imposes constraints that allow for more efficient estimation of the population mean if correct,

Ulrich Müller: umueller@princeton.edu
Andriy Norets: andriy_norets@brown.edu

but under local violations, they generate local biases that can lead to highly erroneous inference about the population mean.

In this paper, we propose to embed a baseline parametric model into a higher dimensional augmented parametric model so that by construction, large sample posteriors are centered at the semiparametrically efficient estimator, and have a variance equal to the semiparametric efficiency bound. Thus, the parameter of interest in the augmented model does not suffer from local biases, for any local misspecification. The augmented model here really is a model, that is, it fully specifies a data generating process (DGP) and the analysis is still fully Bayesian. Many of the desirable features of Bayesian analysis are therefore preserved, such as the likelihood principle, the automatic coherence of multiple Bayes actions, the ability to flexibly incorporate prior knowledge, and accounting for parameter uncertainty in decision and forecasting problems.

There are two natural alternatives to this approach. The first is to make parametric assumptions in the baseline model that directly yield a likelihood that is centered at the semiparametric efficient estimator. For instance, in the example of the location model, this may be achieved by assuming that the data is Gaussian, as the Gaussian MLE is simply the sample mean. However, the misspecification then becomes first order, and the posterior no longer correctly captures data quantiles. Forecasts implied by the posterior thus become quite misleading, for example. Moreover, due to the misspecification, the posterior variance no longer correctly captures sampling uncertainty of the implied estimator in more general models (cf. Müller (2013)). While this can be corrected, such corrections do not lead to a full-information Bayes analysis, and they therefore lack the above mentioned advantages.

The second alternative is to directly employ Bayesian semiparametric modeling. Under high level assumptions, semiparametric Bernstein-von Mises (BVM) theorems state that in such models the marginal posteriors for the finite dimensional parameters behave like classical semiparametrically efficient estimators; see, for example Shen (2002), Bickel and Kleijn (2012), Castillo (2012), Rivoirard and Rousseau (2012), Kato (2013), Castillo and Nickl (2013), and Castillo and Rousseau (2013). However, this direct Bayes semiparametric approach also has potential shortcomings. On the one hand, the assumptions of semiparametric BVM theorems are notoriously difficult to verify. In the context of models used in economics, we are aware of only one example where the assumptions of a semiparametric BVM theorem are known to hold: a partially linear regression with normal homoskedastic errors and a Gaussian process prior on the nonlinear part of the regression, see Bickel and Kleijn (2012). On the other hand MCMC estimation of models with nonparametric priors could be very computationally expensive or even infeasible for higher dimensions or large sample sizes.

For these reasons, the approach suggested here might be a practically appealing approach to robustify Bayesian inference to local misspecification in many settings: The analysis continues to be fully Bayesian, avoids the theoretical pitfalls and practical complications of high-dimensional priors, and allows researchers to continue to work with (potentially locally misspecified) simple parametric models.

The proposed model augmentation consists of a multiplicative factor that involves scores for the baseline model, the efficient scores for an encompassing semiparametric model, and an auxiliary parameter that has the same dimension as the parameter of interest. The augmented model nests the baseline model as a special case when the auxiliary parameter is zero. We develop a Markov Chain Monte Carlo (MCMC) algorithm to estimate the augmented model for a generic baseline model. The algorithm is based on auxiliary latent variables and acceptance sampling, which handle difficult-to-compute normalization constants induced by the augmentation factors, and Hamiltonian Monte Carlo (HMC). The algorithm only requires the following functions as inputs: logarithms of the baseline likelihood and prior and their derivatives, a function that simulates random variables from the baseline model, baseline scores and efficient scores and their derivatives.

The remainder of the paper is organized as follows. Section 2 develops the theoretical results. We discuss the suggested generic MCMC sampling method for the augmented model in Section 3. Section 4 contains illustrations in Weibull and Student's t regression models. Section 5 concludes.

2. MODEL SETUP AND THEORETICAL RESULTS

Subsection 2.1 sets up the notation and standard asymptotic results for the baseline parametric model. In Subsection 2.2, we define local misspecification and show that it leads to a local bias in the estimation of the baseline model. Subsection 2.3 outlines notation and definitions for efficient estimation in a semiparametric model encompassing the baseline model, which will be used in the following subsection to construct an augmentation of the baseline model that avoids the bias under local misspecification. Finally, Theorem 1 in Subsection 2.5 shows that the posterior distribution in the augmented model converges to a normal distribution with mean equal to a semiparametrically efficient estimator and variance equal to the semiparametric efficiency bound, even if the baseline model is locally misspecified.

In this section we heavily rely on the definitions and basic asymptotics results from van der Vaart (1998), especially Chapter 25 on semiparametric models.

2.1. *Baseline model, notation, and standard asymptotics under correct specification*

Suppose the observations $Y_i \in \mathcal{Y}$, $i = 1, \ldots, n$ are independently identically distributed according to distribution $\mathbb{P}_\theta$, where $\theta \in \mathbb{R}^m$. Suppose $\theta = (\gamma, \zeta)$, where $\gamma = \psi(\mathbb{P}_\theta) \in \mathbb{R}^k$ is the parameter of interest and ζ is a nuisance parameter. Let $\dot{\ell}_\theta$ be the score, so that the MLE $\hat{\theta} = (\hat{\gamma}, \hat{\zeta})$ (or, equivalently, the Bayes estimator) under correct specification and regularity conditions satisfies

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta) &= \frac{1}{\sqrt{n}} I_\theta^{-1} \sum_{i=1}^n \dot{\ell}_\theta(Y_i) + o_{\mathbb{P}_\theta}(1) \\ &= \frac{1}{\sqrt{n}} \begin{pmatrix} I_\gamma & I_{\gamma\zeta} \\ I_{\zeta\gamma} & I_\zeta \end{pmatrix}^{-1} \sum_{i=1}^n \begin{pmatrix} \dot{\ell}_\gamma(Y_i) \\ \dot{\ell}_\zeta(Y_i) \end{pmatrix} + o_{\mathbb{P}_\theta}(1) \Rightarrow_\theta \mathcal{N}(0, I_\theta^{-1}),\end{aligned}$$

where $I_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta']$, $I_\gamma = \mathbb{E}_\theta[\dot{\ell}_\gamma \dot{\ell}_\gamma']$, $I_\zeta = \mathbb{E}_\theta[\dot{\ell}_\zeta \dot{\ell}_\zeta']$, and $I_{\zeta\gamma}' = I_{\gamma\zeta} = \mathbb{E}_\theta[\dot{\ell}_\gamma \dot{\ell}_\zeta']$. Thus, with A denoting the first k columns of the $m \times m$ identity matrix,

$$\sqrt{n}(\hat{\gamma} - \gamma) = \frac{1}{\sqrt{n}} A' I_\theta^{-1} \sum_{i=1}^n \dot{\ell}_\theta(Y_i) + o_{\mathbb{P}_\theta}(1) \Rightarrow_\theta \mathcal{N}(0, A' I_\theta^{-1} A)$$

and equivalently, from taking the inverse of the matrix, with $\hat{I}_\gamma = I_\gamma - I_{\gamma\zeta} I_\zeta^{-1} I_{\zeta\gamma}$

$$\begin{aligned}\sqrt{n}(\hat{\gamma} - \gamma) &= \frac{1}{\sqrt{n}} \hat{I}_\gamma^{-1} \sum_{i=1}^n \left(\dot{\ell}_\gamma(Y_i) - I_{\gamma\zeta} I_\zeta^{-1} \dot{\ell}_\zeta(Y_i) \right) + o_{\mathbb{P}_\theta}(1) \qquad (1) \\ &= \frac{1}{\sqrt{n}} \hat{I}_\gamma^{-1} \sum_{i=1}^n \hat{\ell}_\gamma(Y_i) + o_{\mathbb{P}_\theta}(1) \Rightarrow_\theta \mathcal{N}(0, \hat{I}_\gamma^{-1}).\end{aligned}$$

Note that $\hat{\ell}_\gamma$ is the residual of the projection of $\dot{\ell}_\gamma$ on $\dot{\ell}_\zeta$, so $\mathbb{E}_\theta[\hat{\ell}_\gamma(Y_i)\dot{\ell}_\zeta(Y_i)] = 0$.

2.2. *Bias under local misspecification*

Suppose the baseline model $\mathbb{P}_\theta$ is embedded in a semiparametric model $\mathbb{P}_{\theta,\eta}$, where $\eta \in H$ is nonparametric and $\mathbb{P}_{\theta,\eta_0} = \mathbb{P}_\theta$. Importantly, the semiparametric specification here is such the parameter of interest $\psi(\mathbb{P}_{\theta,\eta})$ remains γ, even if $\eta \neq \eta_0$.

Let η_t, $t \in [0,\infty)$ be one dimensional paths through H starting at η_0. By Lemma 25.14 in van der Vaart (1998), under the assumption of differentiability in quadratic mean at $t = 0$, these paths are characterized by their corresponding score g with $\mathbb{E}_\theta[g(Y_i)] = 0$, $\mathbb{E}_\theta[g(Y_i)^2] < \infty$, and

$$\log \prod_{i=1}^n \frac{d\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}}{d\mathbb{P}_\theta}(Y_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(Y_i) - \tfrac{1}{2}\mathbb{E}_\theta[g(Y_i)^2] + o_{\mathbb{P}_\theta}(1). \tag{2}$$

Denote the set of scores that are obtained in this manner by the *tangent set* $\dot{\mathcal{P}}_\theta$ for η. We are exclusively concerned with such local misspecifications of the baseline model, that is, under DGPs where $\eta_t = \eta_{1/\sqrt{n}}$, as in the above equation.

Now for any g, we can characterize the local bias of $\hat{\gamma}$ induced by such local misspecification using contiguity and LeCam's Third Lemma (Example 6.7, page 90 in van der Vaart (1998)). In particular,

$$\left(\sqrt{n}(\hat{\theta} - \theta), \log \prod_{i=1}^n \frac{d\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}}{d\mathbb{P}_\theta}(Y_i)\right)$$
$$\Rightarrow_\theta \mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\mathbb{E}_\theta[g(Y_i)^2] \end{pmatrix}, \begin{pmatrix} I_\theta^{-1} & \cdot \\ \mathbb{E}_\theta[I_\theta^{-1}\dot{\ell}_\theta(Y_i)g(Y_i)] & \mathbb{E}_\theta[g(Y_i)^2] \end{pmatrix}\right),$$

so that under $\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}$,

$$\sqrt{n}(\hat{\theta} - \theta) \Rightarrow_{\theta,\eta_{1/\sqrt{n}}} \mathcal{N}(\mathbb{E}_\theta[I_\theta^{-1}\dot{\ell}_\theta(Y_i)g(Y_i)], I_\theta^{-1}) \tag{3}$$

and

$$\begin{aligned} \sqrt{n}(\hat{\gamma} - \gamma) &\Rightarrow_{\theta,\eta_{1/\sqrt{n}}} \mathcal{N}(\mathbb{E}_\theta[A' I_\theta^{-1}\dot{\ell}_\theta(Y_i)g(Y_i)], A' I_\theta^{-1} A) \\ &\sim \mathcal{N}(\mathbb{E}_\theta[\hat{I}_\gamma^{-1}\hat{\ell}_\gamma(Y_i)g(Y_i)], \hat{I}_\gamma^{-1}). \end{aligned}$$

Thus, unless $\mathbb{E}_\theta[\hat{\ell}_\gamma(Y_i)g(Y_i)] = 0$ for all $g \in \dot{\mathcal{P}}_\theta$, ignoring the misspecification leads to non-zero local biases.

2.3. *Semiparametrically efficient estimation*

Consider paths of the form $t \mapsto \mathbb{P}_{\theta+at,\eta_t}$, as on page 369 in van der Vaart (1998). Then

$$\frac{\partial \log d\mathbb{P}_{\theta+at,\eta_t}}{\partial t}|_{t=0} = a'\dot{\ell}_\theta + g = a'_\gamma\dot{\ell}_\gamma + a'_\zeta\dot{\ell}_\zeta + g$$

and for $\psi(\mathbb{P}_{\theta+at,\eta_t}) = \gamma + a_\gamma t$, we find that $\partial\psi(\mathbb{P}_{\theta+at,\eta_t})/\partial t|_{t=0} = a_\gamma$. So γ is *differentiable as a parameter on the model* if and only if there exists $\tilde{\psi}$ such that

$$a_\gamma = \mathbb{E}_\theta[\tilde{\psi}(Y_i)(a'_\gamma\dot{\ell}_\gamma(Y_i) + a'_\zeta\dot{\ell}_\zeta(Y_i) + g(Y_i))]$$

for any a and $g \in \dot{\mathcal{P}}_\theta$. Setting a_γ to zero, we can see that it is necessary that

$$\mathbb{E}_\theta[\tilde{\psi}(Y_i)\dot{\ell}_\varsigma(Y_i)'] = 0 \text{ and } \mathbb{E}_\theta[\tilde{\psi}(Y_i)g(Y_i)] = 0 \tag{4}$$

for any $g \in \dot{\mathcal{P}}_\theta$. Any semiparametrically efficient estimator T^* of γ has an asymptotically linear representation in terms of this *efficient influence function* $\tilde{\psi}$ (equation (25.22) in van der Vaart (1998))

$$\sqrt{n}(T^* - \gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\psi}(Y_i) + o_{\mathbb{P}_\theta}(1) \Rightarrow \mathcal{N}(0, \mathbb{E}_\theta[\tilde{\psi}(Y_i)\tilde{\psi}(Y_i)']). \tag{5}$$

Furthermore, proceeding as in Lemma 25.25 in van der Vaart (1998), with Π_γ the orthogonal projection operator on the closure of the space of square integrable functions (relative to $\mathbb{P}_\theta$) spanned by linear combinations of $\dot{\mathcal{P}}_\theta$ and elements of ℓ_ς, we have

$$\tilde{\psi} = \tilde{I}_\gamma^{-1}\tilde{\ell}_\gamma \text{ where } \tilde{\ell}_\gamma = \dot{\ell}_\gamma - \Pi_\gamma\dot{\ell}_\gamma \text{ and } \tilde{I}_\gamma = \mathbb{E}_\theta[\tilde{\ell}_\gamma(Y_i)\tilde{\ell}_\gamma(Y_i)']. \tag{6}$$

From this definition of the efficient score $\tilde{\ell}_\gamma$ it follows that

$$\mathbb{E}_\theta[\tilde{\ell}_\gamma(Y_i)\dot{\ell}_\gamma(Y_i)'] = \mathbb{E}_\theta[\tilde{\ell}_\gamma(Y_i)\tilde{\ell}_\gamma(Y_i)'] = \tilde{I}_\gamma. \tag{7}$$

2.4. *Model augmentation*

Now consider an augmented model $\mathbb{Q}_{\theta,\delta}$ with parameters $\theta \in \mathbb{R}^m$ and $\delta \in \mathbb{R}^k$. The augmented model is constructed so that it encompasses the baseline model: $\mathbb{Q}_{\theta,0} = \mathbb{P}_\theta$ and has a score $(\dot{\ell}_\theta', \dot{\ell}_\delta')'$ satisfying

$$\dot{\ell}_\delta(Y_i) = \dot{\ell}_\gamma(Y_i) - \tilde{\ell}_\gamma(Y_i) \tag{8}$$

at θ and $\delta = 0$, that is, the score associated with the additional parameter is the difference between the score for the parameter of interest in the baseline model, and the efficient score.

One explicit construction for $\mathbb{Q}_{\theta,\delta}$ (cf. Example 25.16 in van der Vaart (1998)) is

$$q(y|\theta,\delta) = c(\theta,\delta)k(y,\theta,\delta)p(y|\theta), \tag{9}$$

where $p(y|\theta)$ is the baseline density under $\mathbb{P}_\theta$ relative to ν, $c(\theta,\delta)$ is the normalization constant chosen so that $\int q(y|\theta,\delta)d\nu(y) = 1$, $k(y,\theta,\delta) = k_0(\delta'(\dot{\ell}_\gamma(y) - \tilde{\ell}_\gamma(y)))$, and k_0 is a bounded nonnegative function with $k_0(0) = k_0'(0) = 1$ such as

$$k_0(z) = 2(1 + e^{-2z})^{-1}. \tag{10}$$

Suppose the standard asymptotic expansion for the MLE in the augmented model holds at θ and $\delta = 0$,

$$\sqrt{n}\begin{pmatrix} \hat{\theta}^a - \theta \\ \hat{\delta}^a \end{pmatrix} = \frac{1}{\sqrt{n}} \begin{pmatrix} I_\theta & I_{\theta\delta} \\ I_{\delta\theta} & I_\delta \end{pmatrix}^{-1} \sum_{i=1}^n \begin{pmatrix} \dot{\ell}_\theta(Y_i) \\ \dot{\ell}_\delta(Y_i) \end{pmatrix} + o_{\mathbb{P}_\theta}(1) \Rightarrow_\theta \mathcal{N}\left(0, \begin{pmatrix} I_\theta & I_{\theta\delta} \\ I_{\delta\theta} & I_\delta \end{pmatrix}^{-1}\right). \tag{11}$$

As in the parametric case without augmentation (equation (1)), the resulting expansion of the MLE for γ simply involves the residual variation in the score $\dot{\ell}_\gamma$, after projecting out variation that comes from the nuisance scores $\dot{\ell}_\zeta$ and $\dot{\ell}_\delta$. Note that $I_{\delta\zeta} = \mathbb{E}[\dot{\ell}_\delta \dot{\ell}_\zeta'] = \mathbb{E}[\dot{\ell}_\gamma \dot{\ell}_\zeta'] = I_{\gamma\zeta}$, $I_\delta = I_{\gamma\delta} = I_\gamma - \tilde{I}_\gamma$, and, thus, by a blocked inverse formula,

$$[I_{\gamma\zeta}\; I_{\gamma\delta}] \begin{pmatrix} I_\zeta & I_{\zeta\delta} \\ I_{\delta\zeta} & I_\delta \end{pmatrix}^{-1} = [I_{\gamma\zeta}\; I_\gamma - \tilde{I}_\gamma] \begin{pmatrix} I_\zeta & I_{\zeta\gamma} \\ I_{\gamma\zeta} & I_\gamma - \tilde{I}_\gamma \end{pmatrix}^{-1} = [\mathbb{0}_{k\times m-k} \mathbb{1}_k],$$

where $\mathbb{0}_{k\times m-k}$ is a $k \times m - k$ matrix of zeros and $\mathbb{1}_k$ is a $k \times k$ identity matrix. We thus find that the effective score has variance

$$I_\gamma - [I_{\gamma\zeta}\; I_{\gamma\delta}] \begin{pmatrix} I_\zeta & I_{\zeta\delta} \\ I_{\delta\zeta} & I_\delta \end{pmatrix}^{-1} \begin{bmatrix} I_{\zeta\gamma} \\ I_{\delta\gamma} \end{bmatrix} = \tilde{I}_\gamma$$

as required. Explicitly calculating the effective score yields $\tilde{\ell}_\gamma$, as expected. Thus, the MLE $\hat{\gamma}^a$ for γ in the augmented model satisfies

$$\sqrt{n}(\hat{\gamma}^a - \gamma) = \frac{1}{\sqrt{n}} \tilde{I}_\gamma^{-1} \sum_{i=1}^n \tilde{\ell}_\gamma(Y_i) + o_{\mathbb{P}_\theta}(1). \tag{12}$$

Thus, it is semiparametrically efficient. Note that if (12) holds under $\mathbb{P}_\theta$, then by the definition of contiguity, it also holds under any $\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}$ satisfying (2), so that also

$$\sqrt{n}(\hat{\gamma}^a - \gamma) = \frac{1}{\sqrt{n}} \tilde{I}_\gamma^{-1} \sum_{i=1}^n \tilde{\ell}_\gamma(Y_i) + o_{\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}}(1)$$

and by (4) and (6), $\hat{\gamma}^a$ is asymptotically locally unbiased under local misspecification.

From (6), we have that the asymptotic variance of any efficient estimator T^* satisfies

$$\mathbb{E}_\theta[\tilde{\psi}(Y_i)\tilde{\psi}(Y_i)'] = \tilde{I}_\gamma^{-1}$$

and

$$\tilde{\ell}_\gamma = \mathbb{E}_\theta[\tilde{\psi}(Y_i)\tilde{\psi}(Y_i)']^{-1}\tilde{\psi}. \tag{13}$$

Thus, to obtain $\tilde{\ell}_\gamma$ for the construction of the augmented model it suffices to know the asymptotically linear representation of the semiparametrically efficient estimator T^* of γ.

2.5. *Asymptotics for posterior in augmented model*

Let $Y = \{Y_1, \dots, Y_n\}$ denote a sample of i.i.d. observations and $\Pi(\theta, \delta|Y)$ and $\Pi(\gamma|Y)$ denote the posterior distributions for (θ, δ) and γ that correspond to a prior density $\pi(\theta, \delta)$ and a likelihood function implied by the augmented model $\mathbb{Q}_{\theta,\delta}$.

The following theorem shows that under local misspecification of the baseline model, the posterior for γ in the augmented model has the same asymptotic approximation as the posterior for γ in a Bayesian semiparametric model where a semiparametric BvM (see, for example, Bickel and Kleijn (2012)) holds.

THEOREM 1: *Assume*

(i) in the encompassing semiparametric model, there exists a semiparametrically efficient estimator T^* for γ satisfying (4)-(7)*;*

(ii) the asymptotic expansion of the MLE in the augmented model in (11) *holds;*

(iii) the augmented model is differentiable in quadratic mean at θ and $\delta = 0$ with nonsingular Fisher information matrix and for any $\epsilon > 0$ there exists a sequence of tests ϕ_n satisfying

$$\mathbb{E}_{\mathbb{Q}_{\theta,0}}\phi_n(Y) \to 0, \qquad \sup_{||(\tilde\theta,\delta)-(\theta,0)||>\epsilon} \mathbb{E}_{\mathbb{Q}_{\tilde\theta,\delta}}(1-\phi_n(Y)) \to 0;$$

(iv) the prior density π is positive and continuous at θ and $\delta = 0$.
Then,

$$d_{TV}\left(\Pi(\gamma|Y), \mathcal{N}(\hat\gamma^a, \tfrac{1}{n}\tilde I_\gamma^{-1})\right) = o_{\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}}(1)$$

for smooth paths η_t satisfying (2), where d_{TV} denotes the total variation distance.

PROOF: Under assumptions (iii) and (iv), the Bernstein-von Mises (BvM) theorem (Theorem 10.1 in van der Vaart (1998)) applies under correct specification. The BvM theorem and the discussion on the alternative centering in the BvM on p. 144 in van der Vaart (1998) combined with the asymptotic MLE expansion assumed in (ii) yields

$$d_{TV}\left(\Pi(\theta,\delta|Y), \mathcal{N}\left((\hat\theta^a,\hat\delta^a), \tfrac{1}{n}\begin{pmatrix} I_\theta & I_{\theta\delta} \\ I_{\delta\theta} & I_\delta \end{pmatrix}^{-1}\right)\right) = o_{\mathbb{P}_\theta}(1).$$

Since the total variation distance between two marginal distributions is bounded by the total variation distance between the corresponding joint distributions, it follows that

$$d_{TV}\left(\Pi(\gamma|Y), \mathcal{N}(\hat\gamma^a, \tfrac{1}{n}\tilde I_\gamma^{-1})\right) = o_{\mathbb{P}_\theta}(1).$$

The theorem's claim follows by contiguity of $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}$.

Q.E.D.

Let us briefly comment on elementary sufficient conditions for some of the assumptions in the theorem. It follows from Lemmas 10.3-6 in van der Vaart (1998) that the existence of uniformly consistent tests assumed in (iii) is implied by quadratic mean differentiability, identifiability, and compactness of the parameter space. The asymptotic expansion of the MLE assumed in (ii) is implied by quadratic mean differentiability, nonsingularity of the Fisher information, consistency of the MLE estimator, and some integrability conditions, see Theorem 5.39 in van der Vaart (1998).

3. AUGMENTED POSTERIOR SIMULATION

3.1. *Normalization constants, auxiliary latent variables, and acceptance sampling*

The baseline or original likelihood contribution for observation Y_i is denoted by $p(Y_i|\theta)$. To accommodate models with covariates one could add the covariates in the conditioning set of $p(Y_i|\theta)$; we omit this for notation simplicity. The likelihood contribution of observation Y_i in the augmented model is denoted by $q(Y_i|\theta,\delta)$ defined in (9) and (10) where the augmentation factor $k(Y_i,\theta,\delta)$ has a finite upper bound $\bar k$ and $c(\theta,\delta)$ is a difficult to compute normalization constant. The posterior distribution for the augmented model is given by

$$\pi(\theta,\delta|Y) \propto \prod_{i=1}^{n} q(Y_i|\theta,\delta)\pi(\theta,\delta), \tag{14}$$

where $\pi(\delta)$ and $\pi(\theta)$ are the prior densities. Note that standard MCMC algorithms, such as a Metropolis-Hastings algorithm, do note require the normalization constant $p(Y)$ but would require $c(\theta, \delta)$.

Following Rao, Lin, and Dunson (2016), we use auxiliary latent variables and acceptance sampling to avoid computation of $c(\theta, \delta)$ in the posterior simulator. Let us represent the distribution $q(Y_i|\theta, \delta)$ as if Y_i is obtained by an acceptance sampling algorithm with target density $q(\cdot|\theta, \delta)$, proposal density $p(\cdot|\theta)$, and rejected draws $\tilde{Y}_i = \{\tilde{Y}_{i,j},\ j = 1, \ldots, J_i\}$. In this acceptance sampling algorithm, a proposal $\tilde{Y}_{i,j}$ is simulated from $p(\cdot|\theta)$ and rejected with probability $1 - k(\tilde{Y}_{i,j}, \theta, \delta)/\bar{k}$. For k as in (10), $\bar{k} = 2$, so the rejection probability is not very large. The joint distribution of the accepted draw and the rejected draws can be expressed as follows,

$$\pi(Y_i, \tilde{Y}_i|\theta, \delta) = p(Y_i|\theta)\frac{k(Y_i, \theta, \delta)}{\bar{k}} \cdot \prod_{j=1}^{J_i} p(\tilde{Y}_{i,j}|\theta)\left(1 - \frac{k(\tilde{Y}_{i,j}, \theta, \delta)}{\bar{k}}\right). \tag{15}$$

It is easy to check that the marginal density for Y_i is the target

$$q(Y_i|\theta, \delta) = \sum_{J_i=0}^{\infty} \int \pi(Y_i, \tilde{Y}_i|\theta, \delta) d\tilde{Y}_{i,1} \cdots d\tilde{Y}_{i,J_i}.$$

Therefore, the joint posterior for θ, δ and the auxiliary latent variables $\tilde{Y} = \{\tilde{Y}_i,\ i = 1, \ldots, n\}$,

$$\pi(\theta, \delta, \tilde{Y}|Y) \propto \prod_{i=1}^{n} \pi(Y_i, \tilde{Y}_i|\theta, \delta)\pi(\theta, \delta) \tag{16}$$

implies the marginal posterior of interest $\pi(\theta, \delta|Y)$ in (14) and the draws $(\theta^m, \delta^m, \tilde{Y}^m)$, $m = 1, \ldots, M$ from a Markov chain with stationary distribution (16) can be used to approximate (integrals with respect to) $\pi(\theta, \delta|Y)$.

3.2. *MCMC*

An MCMC algorithm for simulation from (16) consists of two main blocks: (i) $(\theta^m, \delta^m) \sim \pi(\theta, \delta|\tilde{Y}^{m-1}, Y)$ and (ii) $\tilde{Y}^m \sim \pi(\tilde{Y}|\delta^m, \theta^m, Y)$. For the block $\pi(\theta, \delta|\tilde{Y}^{m-1}, Y)$ one could use a Metropolis-Hastings algorithm with a target proportional to (16); in our applications we use HMC for this block as implemented in the Matlab HMC package. To simulate from block $\pi(\tilde{Y}|\delta^m, \theta^m, Y)$ we run the acceptance sampling algorithm described above (15) for each i using (δ^m, θ^m) to obtain the rejected draws $\tilde{Y}_i^m$. The accepted draw can be ignored as it is independent of the rejected draws and the distribution of the rejected draws $\tilde{Y}_i^m$ is proportional to (15) as desired.

The MCMC algorithm is implemented in Matlab for a generic baseline model for which the user needs to supply the following functions: logarithms of the baseline likelihood and prior and their derivatives, a function that simulates Y_i from the baseline model, scores and efficient scores and their derivatives.

4. APPLICATIONS

4.1. *Monte Carlo Simulation for Weibull Regression*

A Weibull regression model, for a random sample of responses y_i and covariate vectors x_i, $i = 1, \dots, n$, is given by

$$p(y_i|x_i,\alpha,\beta) = (\alpha/\lambda)(y_i/\lambda)^{\alpha-1}\exp(-(y_i/\lambda_i)^{\alpha}), \text{ where } \lambda_i = \exp(\beta' x_i))/\Gamma(1+1/\alpha).$$

A typical application of the Weibull regression model involves durations y_i whose baseline hazard function is of the Weibull form, and the individual heterogeneity in durations is modeled by the factor of proportionality $\exp(\beta' x_i)$. The conditional expectation and variance of responses are given by

$$\mathbb{E}(y_i|x_i) = \exp(\beta' x_i) \text{ and } \sigma^2(x_i) = \exp(2\beta' x_i)\cdot[\Gamma(1+2/\alpha)/\Gamma(1+1/\alpha)^2 - 1].$$

The encompassing semiparametric model for the parameter of interest β is defined by the conditional moment restriction $\mathbb{E}(y_i|x_i) = \exp(\beta' x_i)$. In the semiparametric version of the model, no parametric form of the baseline hazard is specified, but the coefficient β continues to have the same interpretation of measuring how the regressors affect the baseline hazard through the factor $\exp(\beta' x_i)$.

Even under local misspecification of the baseline Weibull model, the conditional variance continues to equal to $\sigma^2(x_i)$ to first order, so the semiparametrically efficient estimator is simply the corresponding weighted nonlinear least squares estimator $\hat{\beta}$ which solves

$$\sum_{i=1}^{n} x_i \left[y_i \exp(-\hat{\beta}' x_i) - 1 \right] = 0.$$

In light of (13), the efficient score is thus given by

$$\tilde{\ell}_\beta(y_i) = x_i\,(y_i - \exp(\beta' x_i))\exp(\beta' x_i)/\sigma^2(x_i).$$

In the following Monte Carlo study we illustrate that our suggested model augmentation moves the posteriors of β closer to a normal distribution with mean equal to the semiparametrically efficient estimator and variance equal to the variance of the semiparametrically efficient estimator.

We simulate 100 datasets of size $n = 250$ and $n = 1000$ from the (correctly specified) Weibull regression model with parameters $\alpha = 0.5$, $\beta = (1,1)$, $x_{i1} = 1$, and $x_{i2} \sim \mathcal{N}(0,1)$. The priors for β_k and $\log(\alpha)$ are normal with mean 0 and variance 100. The prior for the augmentation parameter δ is a normal centered at zero; the prior variance is set to an estimate of the asymptotic variance of the MLE for δ under the assumption of no misspecification ($\delta = 0$).

To estimate the baseline model we use the Matlab HMC package. The augmented model is estimated by the MCMC algorithm described in Section 3.2. We found the sampler to mix well; computing times for the augmented model with $n = 250$ are less than a minute on a laptop.

Table I shows the Monte Carlo averages of the difference between the semiparametrically efficient estimator and the posterior mean $\mathbb{E}(\beta|Y)$ and the ratio of the posterior standard deviation to the standard deviation of the semiparametrically efficient estimator in the baseline and augmented models.

As can be seen from the table, in the augmented model, the Bayesian estimator is on average closer to the semiparametrically efficient estimator and the posterior standard deviation is larger and closer to the standard deviation of the semiparametrically efficient estimator, as suggested by our theoretical results.

TABLE I

MONTE CARLO AVERAGES: WEIBULL REGRESSION

	n=250		n=1000	
	Baseline	Augmented	Baseline	Augmented
$\lvert\hat{\beta}_1 - E(\beta_1 \lvert Y)\rvert$	0.03	0.03	0.02	0.01
$\lvert\hat{\beta}_2 - E(\beta_2 \lvert Y)$	0.05	0.03	0.03	0.01
$sd(\beta_1 \lvert Y)/sd(\hat{\beta}_1)$	0.97	1.10	0.96	1.02
$sd(\beta_2 \lvert Y)/sd(\hat{\beta}_2)$	0.90	1.05	0.90	0.99

4.2. *Regression with Student's t errors*

A linear regression model with Student's t errors is recommended for modeling heavy tailed data in most Bayesian econometrics textbooks. It is also prescribed as a tool to introduce individual specific variances in normal linear regression, as a Student's t distribution can be represented as a scale mixture of normal distributions, see, for example, Geweke (2005), Greenberg (2012), Koop (2003), and Geweke (1993). In this model, for a random sample of responses y_i and covariate vectors x_i, $i = 1, \ldots, n$,

$$y_i = x_i'\beta + \epsilon_i, \quad \epsilon_i/\sigma \sim p_s(\cdot), \quad p_s(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

In the application below we treat the regression coefficients β as the parameter of interest and the scale σ and the degrees of freedom ν as nuisance parameters.

In a homoskedastic linear regression model with an unknown distribution of the errors, the ordinary least squares (OLS) estimator is semiparametrically efficient with the efficient score given by

$$\tilde{\ell}_\beta(y_i) = x_i(y_i - x_i'\beta)\frac{1}{\text{var}(\epsilon_i)}.$$

In the Student's t model, $\text{var}(\epsilon_i) = \sigma^2\nu/(\nu - 2)$, and this also holds to first order under local misspecification. We can therefore make this substitution in the efficient score.

If only the first k coefficients in β are of interest, $\gamma = [\mathbb{1}_k \; \mathbb{0}]\beta$, then

$$\tilde{\ell}_\gamma(y_i) = [\mathbb{1}_k \;\; H_{k,k}^{-1}H_{k,-k}]x_i(y_i - x_i'\beta)\frac{1}{\text{var}(\epsilon_i)}, \quad \mathbb{E}(x_i x_i')^{-1} = \begin{pmatrix} H_{k,k} & H_{k,-k} \\ H_{-k,k} & H_{-k,-k} \end{pmatrix}. \tag{17}$$

Figure 1 shows how the augmentation alters the Student's t density under different values of the augmentation parameter for the no regressors case.

4.2.1. *Monte Carlo Simulation*

In this subsection, we present a Monte Carlo study illustrating that the model augmentation moves the posteriors of γ closer to a normal distribution with mean equal to the semiparametrically efficient estimator and variance equal to the variance of the semiparametrically efficient estimator.

We simulate 100 datasets of size $n = 250$ and $n = 1000$ from a Student's t regression with DGP parameters $\sigma = 2$, $\nu = 2.5$, $\beta = (1, 1)$, $x_{i1} = 1$, and $x_{i2} \sim \mathcal{N}(0, 1)$. These values are motivated by the application presented in the next section. The priors for β_1 and $\log(\sigma)$ are normal with mean 0 and variance 100. The prior for $\log(\nu - 2)$ is normal with mean $\log(2)$ and

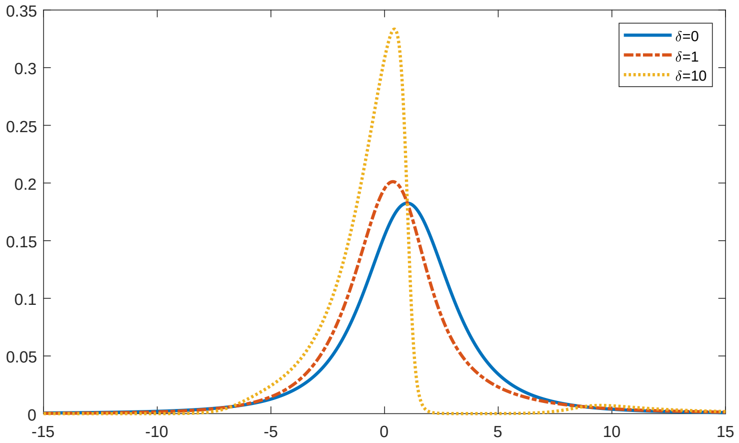

FIGURE 1.—The augmented densities for $x_i = 1$, $\beta = 0$, $\sigma = 2$, $\nu = 2.5$, and $\delta \in \{0, 1, 10\}$.

variance 1. The prior for the augmentation parameter δ is a normal centered at zero; the prior variance is set to an estimate of the asymptotic variance of the MLE for δ under the assumption of no misspecification ($\delta = 0$).

TABLE II

MONTE CARLO AVERAGES: STUDENT'S t REGRESSION

	n=250		n=1000	
	Baseline	Augmented	Baseline	Augmented
$\lvert\hat{\beta}_1 - E(\beta_1\lvert Y)\rvert$	0.17	0.11	0.09	0.04
$\lvert\hat{\beta}_2 - E(\beta_2\lvert Y)$	0.14	0.12	0.09	0.04
$sd(\beta_1\lvert Y)/sd(\hat{\beta}_1)$	0.68	1.08	0.63	1.07
$sd(\beta_2\lvert Y)/sd(\hat{\beta}_2)$	0.72	1.15	0.64	1.08

Table II shows the Monte Carlo averages of the difference between the semiparametrically efficient estimator and the posterior mean $\mathbb{E}(\beta|Y)$ and the ratio of the posterior standard deviation to the standard deviation of the semiparametrically efficient estimator in the baseline and augmented models. As was the case in the Weibull regression case, we again find that the augmentation moves the posteriors towards the semiparametrically efficient estimator and makes the posterior standard deviation larger and closer to the standard deviation of the semiparametrically efficient estimator, as suggested by our theoretical results.

4.2.2. *Application to Incumbency Advantage*

In this application we use data on American congressional elections 1956-1994 to learn about the degree of incumbency advantage, previously analyzed by Jackman (2000) using Bayesian methods. The dataset includes $n = 5090$ observations. The dependent variable is the proportion of the two-party vote won by the Democratic candidate in a district. The covariates include a constant, the proportion of the two-party vote won by the Democratic candidate in the previous election, the previous winning party, indicators for Democratic and Republican incumbency, and 19 dummy variables for time effects. Jackman (2000) argues that for these data, the linear regression errors are heavy tailed (as can be seen in the quantile plots presented in Figure 2 and Jackman (2000)) and that the use of a Student's t distribution is more appropriate. We treat the time effect coefficients as nuisance parameters in the augmented model and use equation (17) for the efficient score. The prior distributions are as in the Monte Carlo simulation. Despite the larger sample size, the sampler still mixes well, see Appendix A for trace plots.

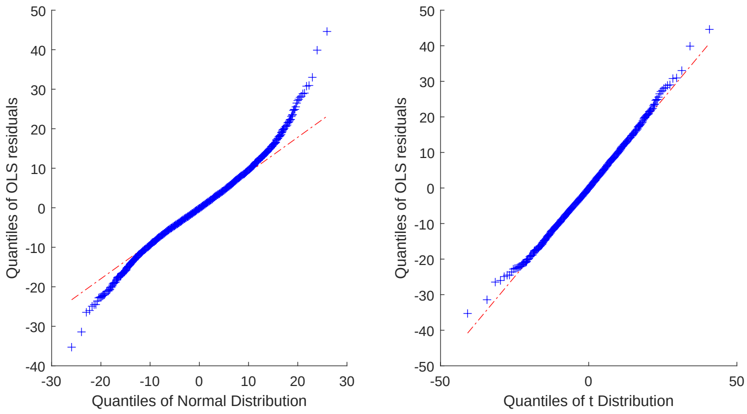

FIGURE 2.—Quantile-quantile plots for OLS residuals, incumbency advantage data.

Figure 3 shows the marginal posterior distributions of the regression coefficients in the baseline and augmented models. Additionally, a normal distribution with mean equal to the OLS estimator and variance equal to the OLS estimator variance is displayed. As can be seen from the figure, the augmented posteriors (solid lines) are moved towards the OLS (dashed lines) relative to the baseline posteriors, as desired.

Figure 4 displays the marginal prior and posterior distributions of the augmentation parameters δ. The posterior strongly prefers non-zero values of δ_i for the first two coordinates, suggesting that the baseline model is misspecified. In fact, the evidence for misspecification here is so overwhelming that it would be sensible to revisit the specification of the baseline model. Such an automatic diagnostic of potential misspecification beyond local deviations accommodated by our theory is another appeal of estimating the augmented model.

As discussed in the introduction, if the sole aim is to obtain semiparametric efficient inference about the regression parameters, one could simply specify the disturbances ϵ_i as Gaussian. But suppose in addition to learning about the degree of incumbency advantage, the researcher uses the estimated model to predict election outcomes. Let us illustrate that the augmented regression model with the Student's t errors can lead to predictions that are substantively different from those generated by the Gaussian linear regression model.

Specifically, consider the probability that a democratic candidate wins in a hypothetical election $n+1$ with the following covariate values: $x_{n+1,2} \in \{31, 33, 35, 37, 39\}$ (vote share of the previous democratic candidate), $x_{n+1,3} = -1$ (previous winner is a Republican), $x_{n+1,4} = 0$ (the democratic candidate is not incumbent), $x_{n+1,5} = 1$ (the republican candidate is incumbent), and $x_{n+1,j} = 0$ for $j = 6, \ldots, 24$. In the Gaussian model, the probability that the democratic candidate wins is computed using a zero mean normal distribution for the regression error and the OLS estimator of the coefficients and the standard deviation of the regression error. In the baseline and augmented Bayesian models, the probability that the democratic candidate wins is the posterior probability that $y_{n+1} > 50$. The probabilities are compared in Table

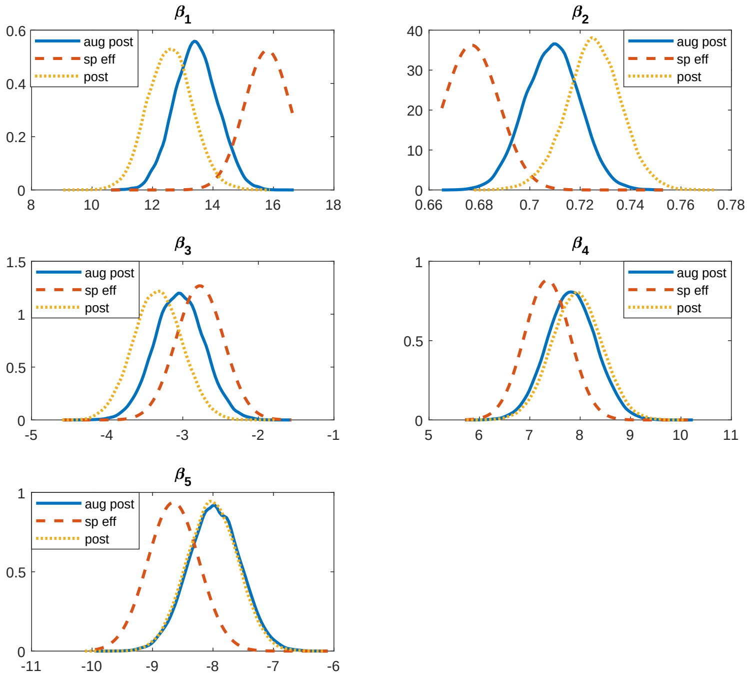

FIGURE 3.—Estimation results for incumbency advantage: posteriors of regression coefficients in the baseline and augmented models; normal distribution centered at the OLS estimator with the corresponding variance.

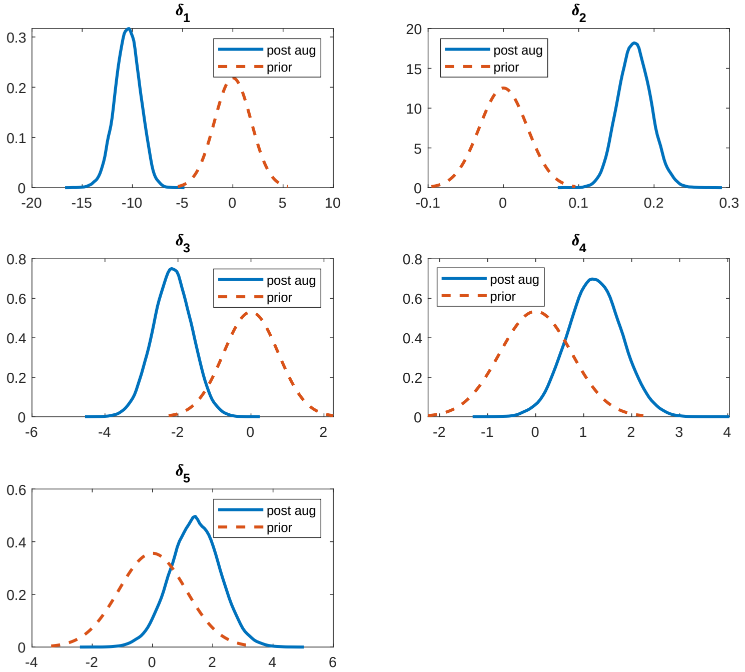

FIGURE 4.—Estimation results for incumbency advantage: marginal priors and posteriors for δ.

TABLE III

PREDICTION FROM GAUSSIAN VS. AUGMENTED t MODEL

Previous dem. vote share	Gaussian, $P(y > 50)$	Augmented model, $P(y > 50)$	Baseline model, $P(y > 50)$
31	0.003	0.010	0.007
33	0.005	0.012	0.010
35	0.009	0.015	0.013
37	0.015	0.020	0.019
39	0.025	0.025	0.026

III. As can be seen from the table, the difference between predictions from the Gaussian linear regression and the augmented Student's t model can be substantial when the predicted probabilities are small, and given the overwhelming evidence for fat tailed disturbances presented in Figure 2, the latter are arguably preferable.

In summary, our augmented model delivers coefficient estimates that are closer to a more robust semi-parametric approach and predictions that rely on a better fitting Student's t distribution for the regression errors.

5. CONCLUSION

In this paper, we propose a method to robustify Bayesian estimation of parametric models. The method applies to settings where the baseline parametric model can be encompassed into a semiparametric model with a known semiparametrically efficient estimator. We augment the baseline likelihood by a multiplicative factor that involves scores for the baseline model, the efficient scores for the encompassing semiparametric model, and an auxiliary parameter that has the same dimension as the parameter of interest. We show that under local misspecification this augmentation asymptotically results in a marginal posterior for the parameter of interest that is normal with mean equal to the semiparametrically efficient estimator and variance equal to the semiparametric efficiency bound; thus, our approach delivers the same asymptotic results as semiparametric BvM theorems, but without the computational and theoretical difficulties inherent in the use of nonparametric priors.

APPENDIX A: AUXILIARY DETAILS FOR APPLICATION

A.1. *Regression with Student's t errors. Incumbency advantage data.*

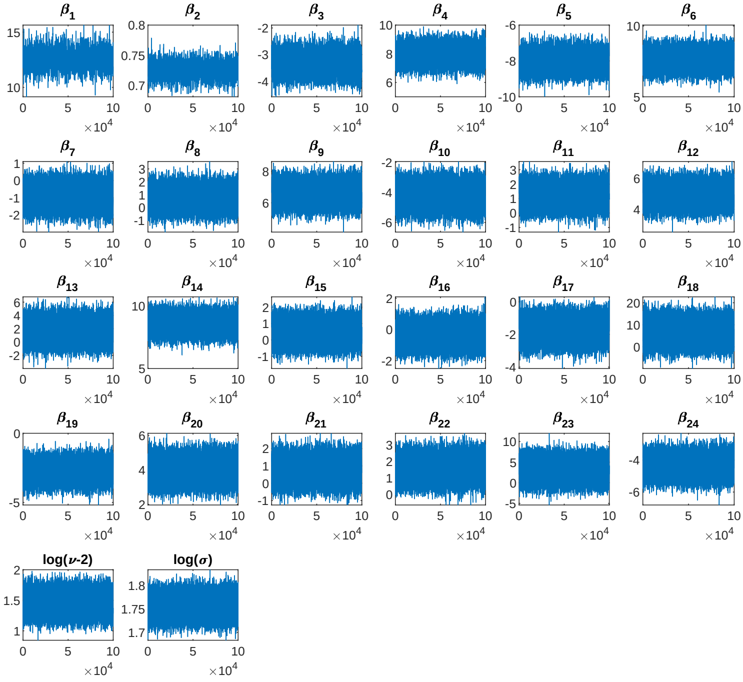

FIGURE A.1.—MCMC trace plots for parameters of a baseline (nonaugmented) regression with Student's t errors estimated on incumbency advantage data.

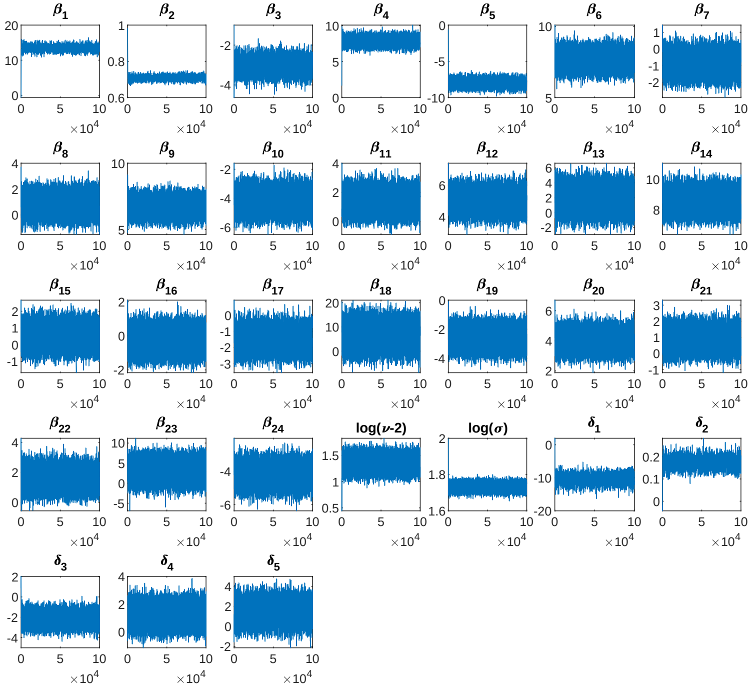

FIGURE A.2.—MCMC trace plots for parameters of an augmented regression with Student's t errors estimated on incumbency advantage data.

REFERENCES

BICKEL, P. J. AND B. J. K. KLEIJN (2012): "The semiparametric Bernstein-von Mises theorem," *Ann. Statist.*, 40 (1), 206–237. [2, 6]

CASTILLO, ISMAEL (2012): "A semiparametric Bernstein-von Mises theorem for Gaussian process priors," *Probability Theory and Related Fields*, 152 (1-2), 53–99. [2]

CASTILLO, ISMAEL AND RICHARD NICKL (2013): "Nonparametric Bernstein-von Mises theorems in Gaussian white noise," *The Annals of Statistics*, 41 (4), 1999–2028. [2]

CASTILLO, ISMAEL AND JUDITH ROUSSEAU (2013): "A General Bernstein–von Mises Theorem in semiparametric models," ArXiv:1305.4482. [2]

GEWEKE, J. (1993): "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics*, 8, S19–S40. [10]

GEWEKE, JOHN (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience. [10]

GREENBERG, EDWARD (2012): *Introduction to Bayesian Econometrics*, Cambridge University Press, 2 ed. [10]

JACKMAN, SIMON (2000): "Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation," *Political Analysis*, 8 (4), 307–332. [11, 12]

KATO, KENGO (2013): "Quasi-Bayesian analysis of nonparametric instrumental variables models," *The Annals of Statistics*, 41 (5), 2359–2390. [2]

KOOP, GARY (2003): *Bayesian econometrics*, Chichester Hoboken, N.J: J. Wiley. [10]

MÜLLER, ULRICH K. (2013): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix," *Econometrica*, 81 (5), 1805–1849. [2]

RAO, VINAYAK, LIZHEN LIN, AND DAVID B. DUNSON (2016): "Data augmentation for models based on rejection sampling," *Biometrika*, 103 (2), 319–335. [8]

RIVOIRARD, V. AND J. ROUSSEAU (2012): "Bernstein-von Mises theorem for linear functionals of the density," *Annals of Statistics*, 40 (3), 1489–1523. [2]

SHEN, XIAOTONG (2002): “Asymptotic Normality of Semiparametric and Nonparametric Posterior Distributions,” *Journal of the American Statistical Association*, 97 (457), 222–235. [2]

VAN DER VAART, A.W. (1998): *Asymptotic Statistics*, Cambridge University Press. [3, 4, 5, 7]

Co-editor [Name Surname; will be inserted later] handled this manuscript.