Submitted to

LOCALLY ROBUST EFFICIENT BAYESIAN INFERENCE

MARCO DEL NEGRO
Macroeconomic and Monetary Studies, Federal Reserve Bank of New York

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 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 augmentation asymptotically results in a marginal posterior for the parameter of interest that is normal with the mean equal to the semiparametrically efficient estimator and the variance equal to the semiparametric efficiency bound. The augmented model nests the baseline model as a special case when the auxiliary parameter is zero. The approach should be especially useful when not only the parameters but other aspects of the distribution are of interest. We develop an MCMC algorithm for the augmented model estimation. The approach is illustrated 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 high probability, 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 first-order correct.

Marco Del Negro: marco.delnegro@ny.frb.org
Ulrich Müller: umueller@princeton.edu
Andriy Norets: andriy_norets@brown.edu

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, but under local violations, they generate local biases that lead to potentially 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.

A natural alternative to our approach is to directly employ Bayesian semiparametric modelling. 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 has some considerable shortcomings. First, 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 have been verified: 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). Second, it is possible to construct examples based on sieve priors for which such semiparametric BVM theorems do not hold, see Appendix B.1. Finally, 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 alternative in many settings. The proposed model augmentation consists of 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. 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.

We illustrate our approach in a linear regression with Student's t errors in Section 4; work on illustrations in several other models is currently underway.

2. MODEL AUGMENTATION

In this section we rely heavily 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*

For simplicity, consider the case where 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 a positive and continuous prior density) satisfies under correct specification that

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

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 $\hat{I}_\gamma = I_\gamma - I_{\gamma\zeta} I_\zeta^{-1} I_{\zeta\gamma}$

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

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

2.2. *Bias under local misspecification*

We consider misspecifications of the baseline model of a nonparametric form: Let $\mathbb{P}_{\theta,\eta}$ with $\eta \in H$ nonparametric be the distribution of the observations, where $\mathbb{P}_{\theta,\eta_0} = \mathbb{P}_\theta$. Let η_t, $t \in [0, \infty)$ be one dimensional paths through H starting at η_0. Under the regularity conditions in chapter 25.3 of van der Vaart (1998), (the appropriate subset of) these paths are characterized by their corresponding score g, as in

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

(the first displayed equation on page 363 in van der Vaart (1998)). Denote the set of scores that are obtained in this manner by the *tangent set* $\dot{\mathcal{P}}_\theta$. 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 $\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}$,

$$\sqrt{n}(\hat{\theta}-\theta) \Rightarrow_{\theta,\eta_{1/\sqrt{n}}} \mathcal{N}(\mathbb{E}_\theta[I_\theta^{-1}\dot{\ell}_\theta(Y_i)g(Y_i)], I_\theta^{-1}) \tag{2}$$

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 and model augmentation*

Now 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 part of the model if and only if there exists $\tilde{\psi} \in \dot{\mathcal{P}}_\theta$ such that

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

and setting a_γ to zero, we can see that it is necessary that

$$0 = \mathbb{E}_\theta[\tilde{\psi}(Y_i)\dot{\ell}_\zeta(Y_i)] = \mathbb{E}[\tilde{\psi}(Y_i)g(Y_i)] \tag{3}$$

for $g \in \dot{\mathcal{P}}_\theta$, and any semiparametrically efficient estimators T^* of θ satisfies (cf. equation (25.22))

$$\sqrt{n}(T^*-\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\psi}(Y_i) + o_{\mathbb{P}_\theta}(1) \Rightarrow \mathcal{N}(0, \mathbb{E}_\theta[\tilde{\psi}(Y_i)\tilde{\psi}(Y_i)']). \tag{4}$$

Furthermore, proceeding as in Lemma 25.25, with Π_γ the orthogonal projection operator on the linear closure of $\text{lin}\dot{\ell}_\zeta + \dot{\mathcal{P}}_\theta$, we have

$$\tilde{\psi} = \tilde{I}_\gamma^{-1}\tilde{\ell}_\gamma \text{ where } \tilde{\ell}_\gamma = \dot{\ell}_\gamma - \Pi_\gamma\dot{\ell}_\gamma \text{ and } \tilde{I}_\gamma = \mathbb{E}_\theta[\tilde{\ell}_\gamma(Y_i)\tilde{\ell}_\gamma(Y_i)']. \tag{5}$$

From this definition of $\tilde{\ell}_\gamma$ it follows (cf. the proof of Lemma 25.25)

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

Now consider an *augmented* baseline model $\mathbb{Q}_{\theta,\delta}$ with parameters $\theta \in \mathbb{R}^m$ and $\delta \in \mathbb{R}^k$, which is constructed such that

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

where the score s_γ equals

$$s_\gamma(Y_i) = \dot{\ell}_\gamma(Y_i) - \tilde{\ell}_\gamma(Y_i).$$

As in the parametric case above, 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 s_γ. We thus find that the effective score has variance

$$I_\gamma - \begin{pmatrix} I_{\gamma\zeta} \\ I_\gamma - \tilde{I}_\gamma \end{pmatrix}' \begin{pmatrix} I_\zeta & I_{\gamma\zeta} \\ I_{\zeta\gamma} & I_\gamma - \tilde{I}_\gamma \end{pmatrix}^{-1} \begin{pmatrix} I_{\gamma\zeta} \\ I_\gamma - \tilde{I}_\gamma \end{pmatrix} = \tilde{I}_\gamma$$

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

$$\sqrt{n}(\hat{\gamma}^a - \gamma) = \frac{1}{\sqrt{n}} \tilde{I}_\gamma^{-1} \sum_{i=1}^n \tilde{\ell}_\gamma(Y_i) + o_{\mathbb{P}_\theta}(1) \tag{6}$$

so it is semiparametrically efficient. Note that if (6) holds under $\mathbb{P}_\theta$, then by the definition of contiguity, it also holds under any $\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}$ satisfying (1), 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 (3), $\hat{\gamma}^a$ is asymptotically locally unbiased under local misspecification.

The key regularity condition for these results are the assumptions of Lemma 25.25 in van der Vaart (1998).

From (5), we have that the asymptotic variance of any efficient estimator T^* satisfies

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

Thus

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

so all we need to be able to obtain $\tilde{\ell}_\gamma$ for the construction of the augmented model is knowledge of the asymptotically linear representation of the semiparametrically efficient estimator T^* of γ.

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

$$q(y|\theta,\delta) = c(\theta,\delta)k(y,\theta,\delta)p(y|\theta), \tag{7}$$

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 nonnegative function with $k_0(0)=k_0'(0)=1$ such as a bounded above $k_0(z)=2(1+e^{-2z})^{-1}$. For the MCMC algorithm presented in Section 3 below, it is convenient to use a function k_0 that is both bounded above and bounded away from zero, namely

$$k_0(z)=1/2+(1+e^{-4z})^{-1}. \tag{8}$$

The following alternative construction avoids having to explicitly obtain the scores $\dot{\ell}_\gamma$ of the baseline model. In particular, note that

$$\tilde{q}(y|\theta,\delta)=\tilde{c}(\theta,\delta)\,k_0(-\delta'\tilde{\ell}_\gamma(y))\,p(y|(\gamma+\delta,\zeta)) \tag{9}$$

with $\tilde{c}(\theta,\delta)$ chosen such that $\int \tilde{q}(y|\theta,\delta)d\nu(y)=1$ also has the required local properties for small values of $||\delta||$. This follows, since

$$\begin{aligned}
\frac{\partial \log \tilde{q}(y|\theta,\delta)}{\partial\gamma}|_{\delta=0}&=\frac{\partial \log \tilde{c}(\theta,\delta)}{\partial\gamma}|_{\delta=0}+\dot{\ell}_\gamma(y)\\
\frac{\partial \log \tilde{q}(y|\theta,\delta)}{\partial\zeta}|_{\delta=0}&=\frac{\partial \log \tilde{c}(\theta,\delta)}{\partial\zeta}|_{\delta=0}+\dot{\ell}_\zeta(y)\\
\frac{\partial \log \tilde{q}(y|\theta,\delta)}{\partial\delta}|_{\delta=0}&=\frac{\partial \log \tilde{c}(\theta,\delta)}{\partial\delta}|_{\delta=0}-\tilde{\ell}_\gamma(y)+\dot{\ell}_\gamma(y).
\end{aligned}$$

Taking the expectations of these scores at θ and $\delta=0$ and using $\mathbb{E}_\theta[\tilde{\ell}_\gamma(Y_i)]=\mathbb{E}_\theta[\dot{\ell}_\gamma(Y_i)]=0$, we get $\frac{\partial \ln c_{\theta,\delta}}{\partial\theta}|_{\delta=0}=0$ and $\frac{\partial \ln c_{\theta,\delta}}{\partial\delta}|_{\delta=0}=0$. The required second order properties (expectation of the Hessians) follow from the information matrix equality.

3. AUGMENTED POSTERIOR SIMULATION

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

Let $Y=\{y_1,\ldots,y_n\}$ denote a sample of iid observations. 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 (7) and (8) 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

$$\pi(\theta,\delta|Y)\propto\prod_{i=1}^{n}q(y_i|\theta,\delta)\pi(\theta)\pi(\delta), \tag{10}$$

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 the approach from Rao, Lin, and Dunson (2016), we use auxiliary latent variables and acceptance sampling to avoid the computation of $c(\theta,\delta)$ in a posterior simulator. Let us represent the distribution $q(y_i|\theta,\delta)$ as if y_i is obtained by an acceptance sampling algorithm with the target density $q(\cdot|\theta,\delta)$, the source density $p(\cdot|\theta)$, and rejected draws $\tilde{y}_i=\{\tilde{y}_{i,j},\ j=$

$1,\dots,J_i\}$. In this acceptance sampling algorithm, a proposal $\tilde{y}_{i,j}$ is simulated from $p(\cdot|\theta)$ and rejected with probability $1-k(\tilde{y}_{i,j},\theta,\delta)/\bar{k}$. The joint distribution of the accepted draw and the rejected draws can be expressed as follows,

$$\pi(y_i,\tilde{y}_i|\theta,\delta)=p(y_i|\theta)\frac{k(y_i,\theta,\delta)}{\bar{k}}\cdot\prod_{j=1}^{J_i}p(\tilde{y}_{i,j}|\theta)\left(1-\frac{k(\tilde{y}_{i,j},\theta,\delta)}{\bar{k}}\right). \tag{11}$$

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

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

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

$$\pi(\theta,\delta,\tilde{Y}|Y)\propto\prod_{i=1}^{n}\pi(y_i,\tilde{y}_i|\theta,\delta)\pi(\theta)\pi(\delta) \tag{12}$$

implies the marginal posterior of interest $\pi(\theta,\delta|Y)$ in (10) and the draws $(\theta^m,\delta^m,\tilde{Y}^m)$, $m=1,\dots,M$ from a Markov chain with the stationary distribution in (12) can be used to approximate (integrals with respect to) $\pi(\theta,\delta|Y)$.

3.2. *MCMC*

An MCMC algorithm for simulation from (12) consists of two main blocks: (1) $(\theta^m,\delta^m)\sim\pi(\theta,\delta|\tilde{Y}^{m-1},Y)$ and (2) $\tilde{Y}^m\sim\pi(\tilde{Y}|\delta^m,\theta^m,Y)$. For the block $\pi(\theta,\delta|\tilde{Y}^{m-1},Y)$ one could use a Metropolis-Hastings algorithm with a target proportional to (12); in our applications we use HMC as implemented in a Matlab package. To simulate from block $\pi(\tilde{Y}|\delta^m,\theta^m,Y)$ one could run the acceptance sampling algorithm described above (11) 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 (11) 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.

3.3. *Importance sampling*

Importance sampling is an alternative to the MCMC algorithm from Section 3.2 that could be easier to implement if draws θ^m, $m=1,\dots,M$ from the posterior of the baseline model

$$\pi(\theta|Y)\propto\prod_{i=1}^{n}p(y_i|\theta)\pi(\theta) \tag{13}$$

are readily available. Specifically, consider the following importance sampling source distribution for $(\theta,\delta;\tilde{Y},\tilde{y}_{1,J_1+1},\dots,\tilde{y}_{n,J_n+1})$

$$\pi(\theta|Y)\pi(\delta)\prod_{i=1}^{n}\pi(\tilde{y}_{i,J_i+1},\tilde{y}_i|\theta,\delta).$$

Note that we include the accepted draws $\tilde{y}_{i,J_i+1}$ here because without them the intractable constants $c(\theta,\delta)$ would be present in the marginal distribution of the rejected draws and we would need to evaluate them in the computation of the importance sampling weights. The target distribution is

$$\pi(\theta,\delta,\tilde{Y}|Y)\prod_{i=1}^{n}p(\tilde{y}_{i,J_i+1}|\theta),$$

where the target density for $\tilde{y}_{i,J_i+1}$ is taken to be the baseline likelihood (in principle, it could be arbitrary). Then, for a sample from the source density (or MCMC for the source density), $(\theta^m,\delta^m,\tilde{Y}^m,\tilde{y}^m_{1,J_1+1},\ldots,\tilde{y}^m_{n,J_n+1})$, $m=1,\ldots,M$, the importance sampling weights are as follows,

$$w^m \propto \frac{\pi(\theta^m,\delta^m,\tilde{Y}^m|Y)\prod_{i=1}^{n}p(\tilde{y}^m_{i,J_i+1}|\theta^m)}{\pi(\theta^m|Y)\pi(\delta^m)\prod_{i=1}^{n}\pi(\tilde{y}^m_{i,J_i+1},\tilde{y}^m_i|\theta^m,\delta^m)} \propto \prod_{i=1}^{n}\frac{k(y_i,\theta^m,\delta^m)}{k(\tilde{y}^m_{i,J_i+1},\theta^m,\delta^m)}.$$

The last expression in the above display can be normalized and used as importance sampling weights.

4. APPLICATIONS

4.1. *Regression with Student's t errors*

A linear regression model with Student's t errors is recommended for modelling 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 a 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

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

4.1.1. *House price data*

Koop (2003) used data on house prices and covariates from Anglin and Gencay (1996) to illustrate regression with Student's t errors. The dataset includes 546 observations. The dependent variable is the sale price of a house standardized to have sample mean zero and variance one. The covariates are the constant, the lot size of a property (standardized), the number of bedrooms, the number of full bathrooms, and the number of stories excluding basement. We fix

the degrees of freedom parameter to $\nu = 4$ (the posterior mean from Koop (2003)). The prior distributions for β_i, $i = 1, \ldots, 5$ and $\log(\sigma)$ are normal with mean 0 and variance 100.

The prior for the augmentation parameters δ is a multivariate normal centered at zero. The prior variance covariance is set to an estimate of the asymptotic variance of the MLE for δ under the assumption of no misspecification ($\delta_i = 0$, $i = 1, \ldots, 5$ in the data generating process) multiplied by 2.

To estimate the baseline model we use a Matlab's HMC package. The augmented model is estimated by the MCMC algorithm described in Section 3.2. The MCMC algorithms converge quickly as can be assessed in Figures A.1 and A.2 in Appendix A.1 displaying MCMC trace plots.

Figure 1 shows the marginal posterior distributions of the regression coefficients in the baseline and augmented models. Additionally, normal distributions centered at the MLE and OLS with the corresponding estimator variances are displayed. As expected from the standard BVM result, the MLE dash dotted lines are aligned with the dotted lines of the baseline posteriors. The solid augmented posteriors are aligned or at least moved towards the dashed OLS lines as expected from the theoretical results in Section 2.

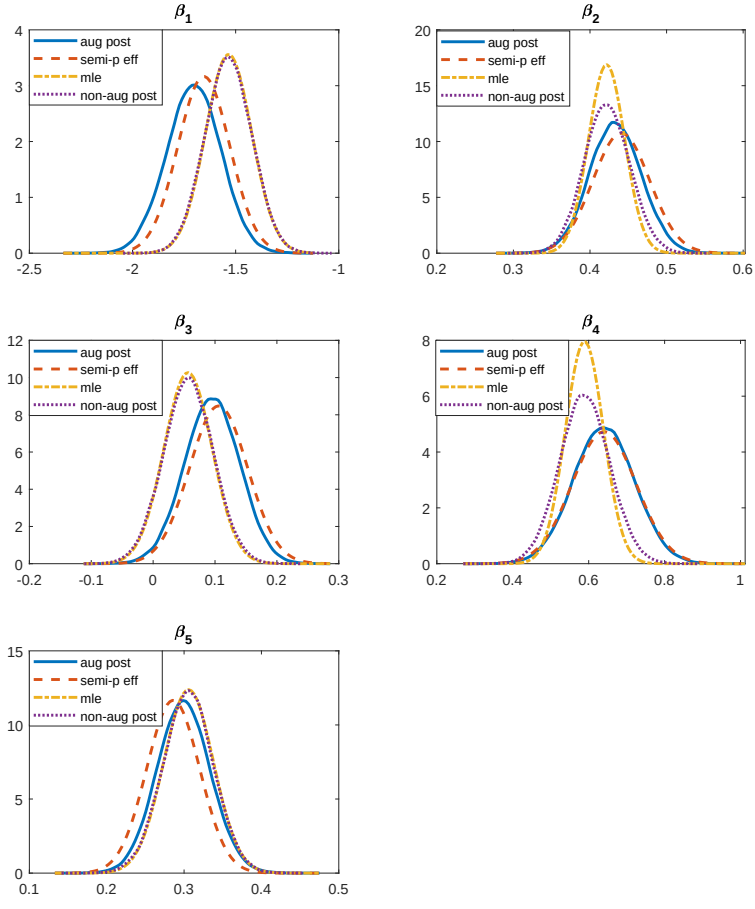

FIGURE 1.—Estimation results for house price data: posteriors of regression coefficients in the baseline and augmented models; normal distributions centered at the MLE and OLS with the corresponding estimator variances.

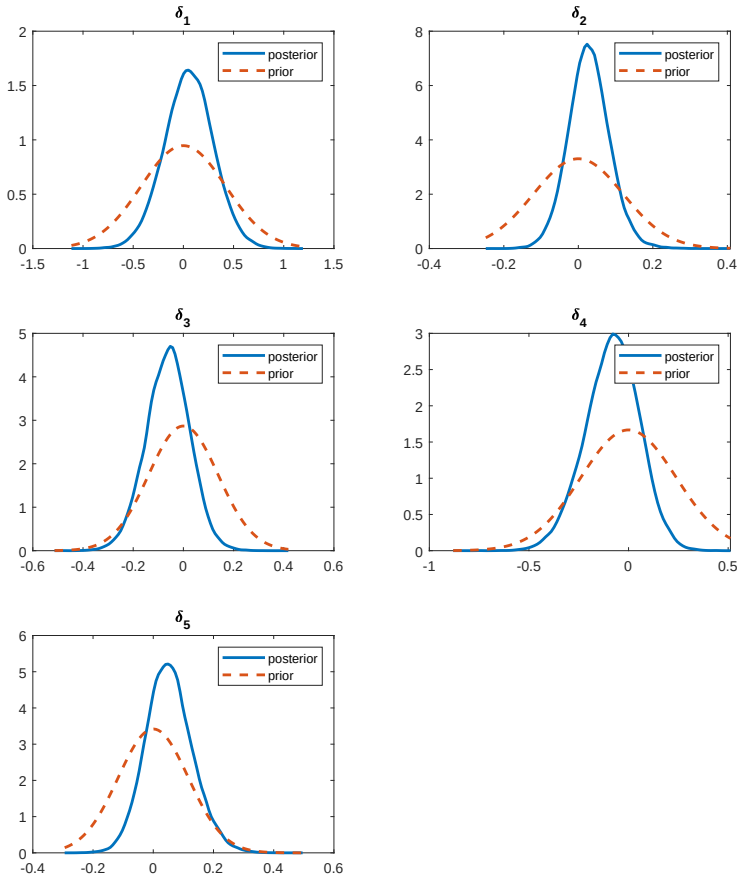

FIGURE 2.—Estimation results for house price data: marginal priors and posteriors for δ.

Figure 2 displays the marginal prior and posterior distributions of the augmentation parameters δ. Zero values of the augmentation parameters are well within the support of the posteriors. These results do not contradict the local misspecification assumption justifying the asymptotic properties of the augmented posteriors.

5. FUTURE WORK

We plan to further illustrate the proposed methodology for robustifying Bayesian inference in parametric models in a number of applications: time series models, models with stochastic volatility, instrumental variable models, Weibull regressions, and others.

APPENDIX A: AUXILIARY DETAILS FOR APPLICATIONS

A.1. *Regression with Student's t errors. House price data.*

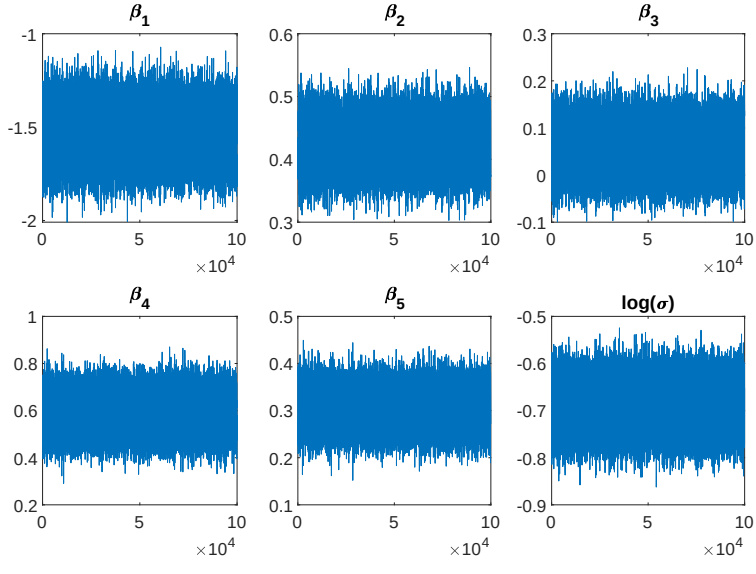

FIGURE A.1.—MCMC trace plots for parameters of a baseline (unaugmented) regression with Student's t errors estimated on house price data.

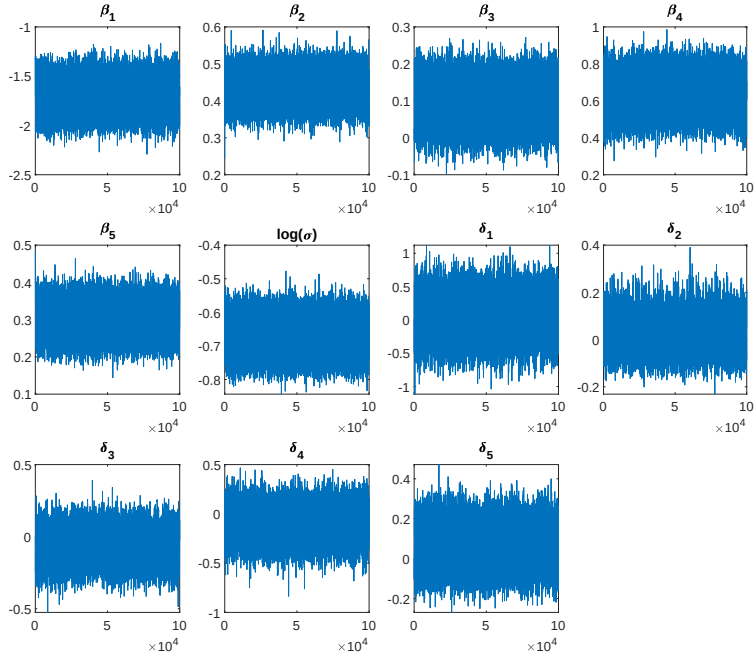

FIGURE A.2.—MCMC trace plots for parameters of an augmented regression with Student's t errors estimated on house price data.

APPENDIX B: AUXILIARY RESULTS

B.1. *Failure of semiparametric BVM theorem under sieve priors*

Consider the problem of estimating a function g on the unit interval based on n equi-spaced observations

$$y_i = g\left(\frac{i-1/2}{n}\right) + \varepsilon_i, \varepsilon_i \sim iid\mathcal{N}(0,1).$$

Let $\hat{\beta}_0 = n^{-1}\sum_{i=1}^n y_i$, $\beta_0 = n^{-1}\sum_{i=1}^n g\left(\frac{i-1/2}{n}\right)$,

$$\hat{\beta}_j = n^{-1}\sum_{i=1}^n \sqrt{2}\cos\left(\pi j\frac{i-1/2}{n}\right) y_i, j = 1,\ldots,n-1$$

$$\beta_j = n^{-1}\sum_{i=1}^n \sqrt{2}\cos\left(\pi j\frac{i-1/2}{n}\right) g\left(\frac{i-1/2}{n}\right), j = 1,\ldots,n-1.$$

By the orthonormality of the type II discrete cosine transform, we then have

$$\hat{\beta}_j \sim \text{ independent } \mathcal{N}(\beta_j, 1/n), j = 0,\ldots,n-1.$$

By standard calculations, a continuous function g will lead to $\beta_j \asymp j^{-2}$, a differentiable function to $\beta_j \asymp j^{-3}$, an infinitely differentiable function to an exponential decay, etc.

Let $\gamma = \sum_{j=1}^{n} \alpha_j \beta_j \in \mathbb{R}$ be the functional of interest, where we normalize $\sum_{j=1}^{n} \alpha_j^2 = 1$.The MLE is

$$\hat{\gamma} = \sum_{j=1}^{n} \alpha_j \hat{\beta}_j \sim \mathcal{N}(\gamma, n^{-1}).$$

B.1.1. *Gaussian Process Prior*

Consider the prior, $\beta_j \sim$ independent $\mathcal{N}(0, c_j)$, $j = 0, \ldots, n-1$ for some $c_j \to 0$ (such as $c_j = 2^{-j}$ or $c_j = j^{-5}$). Under this prior, the posterior is

$$\beta_j | Y \sim \text{ independent } \mathcal{N}\left(\frac{nc_j}{1+nc_j} \hat{\beta}_j, (c_j^{-1} + n)^{-1} \right)$$

where $Y = (y_1, \ldots, y_n)$. For any fixed j, clearly $(c_j^{-1} + n)^{-1} \to n^{-1}$, so the posterior variance of γ will also converge to n^{-1}, just like the MLE. So under a Gaussian process prior, we obtain semiparametrically efficient inference about γ.

B.1.2. *Sieve Prior*

Let the model indexed by the integer $m = 1, \ldots, n$ be such that $\beta_j = 0$ for $j \geq m$, and the prior on β_j, $j < m$ is $\beta_j \sim \mathcal{N}(0, 1)$. Let the prior on the model m be 2^{-m} (and to ensure adding up, let the prior on model $m = n$ be $2 \cdot 2^{-n}$).

Fix $m_0 > 1$. Consider the case where $g = 0$, and $\gamma = \beta_{m_0} = 0$. Then surely, the posterior probability of $m \leq m_0$ is smaller than one for all n—the posterior for m will visit $m > m_0$, but with very low probability. Thus, the posterior variance for γ is a mixture of zero and $(1+n)^{-1}$, with most of the mass on 0, so it is certainly smaller than n^{-1} even as $n \to \infty$.

By contiguity, this continues to hold for local alternatives to $g = 0$ (say, $\beta_0 \neq 0$, $\beta_j = b_j n^{-1/2}$ for $j = 1, \ldots, 2m_0$), where the Bayesian posterior for γ is biased (since it's a mixture between zero with high probability, and $\frac{n}{1+n}\hat{\beta}_{m_0}$). So Bayesian semiparametric inference here will *not* yield semiparametrically efficient inference about γ.

The same applies under a "sieve" Gaussian process prior where the prior $\beta_j \sim \mathcal{N}(0, 1)$ is replaced by $\beta_j \sim \mathcal{N}(0, c_j)$.

One might argue that the failure is non-generic, since the above argument depends on β_j to be very nearly zero for all but a finite number of j. So consider now a true model where $\beta_j = \exp(-\delta_0 j)$, for some $\delta_0 > 0$. (So the truth is very smooth). Consider a prior $\beta_j \sim \mathcal{N}(0, c_j)$, and an exponential prior on m.

The likelihood ratio of $\hat{\beta}_j \sim \mathcal{N}(0, c_j + n^{-1})$ (including the jth component) and $\hat{\beta}_j \sim \mathcal{N}(0, n^{-1})$ (excluding it) is

$$\begin{aligned}
&\frac{\sqrt{n^{-1}}}{\sqrt{c_j + n^{-1}}} \exp[-\tfrac{1}{2}\hat{\beta}_j^2 (\frac{1}{c_j + n^{-1}} - \frac{1}{n^{-1}})] \\
&= \frac{1}{\sqrt{1 + c_j n}} \exp[\tfrac{1}{2}\hat{\beta}_j^2 \frac{c_j n^2}{1 + c_j n}] \\
&\approx \frac{1}{\sqrt{1 + c_j n}} \exp[\tfrac{1}{2}\exp(-2\delta_0 j) \frac{c_j n^2}{1 + c_j n}].
\end{aligned}$$

With $c_j = \exp(-\delta_\pi j)$, the log-likelihood ratio becomes

$$-\tfrac{1}{2}\log(1+\exp(-\delta_\pi j)n) + \tfrac{1}{2}\exp(-2\delta_0 j)\frac{\exp(-\delta_\pi j)n^2}{1+\exp(-\delta_\pi j)n}$$

which is $O(1)$ when $j \approx \frac{2}{2\delta_0+\delta_\pi}\log n$ for $\delta_\pi \geq 2\delta_0$ or $j \approx \frac{1}{\delta_\pi}\log n$ for $\delta_\pi < 2\delta_0$. Initially focus on the latter which covers the case where the prior decay is the same as the decay of the truth (so the truth is as smooth as expected a priori). In this model the posterior for m concentrates around values $m \approx \frac{1}{\delta_\pi}\log n$. The approximate bias of γ induced by shrinking all β_j with $j > m$ to zero thus is

$$\text{bias} = \sum_{j=\frac{1}{\delta_\pi}\log n}^{n} \alpha_j\beta_j = \sum_{j=\frac{1}{\delta_\pi}\log n}^{n} \alpha_j \exp(-\delta_0 j)$$

For $\alpha_j = \exp(-\delta_\gamma j)$, this evaluates to

$$\text{bias} \approx n^{-(\delta_\gamma+\delta_0)/\delta_\pi}$$

so the bias is of order at least $n^{-1/2}$ as long as $\delta_\gamma \leq \frac{1}{2}(\delta_\pi - 2\delta_0)$. This is impossible given the above restriction, so at least in this case, there won't be bias of order $n^{-1/2}$.

So now consider the case where $\delta_\pi \geq 2\delta_0$ (so the truth is less smooth than expected, but there is still exponential decay). Proceeding as above yields

$$\text{bias} \approx n^{-2(\delta_\gamma+\delta_0)/(2\delta_0+\delta_\pi)}$$

so the bias is of order at least $n^{-1/2}$ as long as $\delta_\gamma \leq \frac{1}{4}(\delta_\pi - 2\delta_0)$. So even though everything here is very smooth in the sense of exponentially decaying, there is a systematic bias of order at least $n^{-1/2}$.

REFERENCES

ANGLIN, PAUL M. AND RAMAZAN GENCAY (1996): "Semiparametric estimation of a hedonic price function," *Journal of Applied Econometrics*, 11 (6), 633–648. [8]

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

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. [8]

GEWEKE, JOHN (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience. [8]

GREENBERG, EDWARD (2012): *Introduction to Bayesian Econometrics*, Cambridge University Press, 2 ed. [8]

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. [8, 9]

RAO, VINAYAK, LIZHEN LIN, AND DAVID B. DUNSON (2016): "Data augmentation for models based on rejection sampling," *Biometrika*, 103 (2), 319–335. [6]

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]

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