# BIAS REDUCTION FOR BAYESIAN AND FREQUENTIST ESTIMATORS

C. ALAN BESTER AND CHRISTIAN HANSEN†

ABSTRACT. We show that in parametric likelihood models the first order bias in the posterior mode and the posterior mean can be removed using objective Bayesian priors. These bias-reducing priors are defined as the solution to a set of differential equations which may not be available in closed form. We provide a simple and tractable data dependent prior that solves the differential equations asymptotically and removes the first order bias. When we consider the posterior mode, this approach can be interpreted as penalized maximum likelihood in a frequentist setting. We illustrate the construction and use of the bias-reducing priors in simple examples and a simulation study.

*Keywords: Bias, Objective Bayes, Penalized likelihood*

*JEL Codes: C11, C13*

## 1. INTRODUCTION

In parametric likelihood problems, the maximum likelihood estimator (MLE) has bias that declines with the sample size or total Fisher information. In practice, this bias is often small relative to the standard errors and is typically ignored. However, in many settings, including time series and panel data applications, bias may be considerable in finite samples. In these cases, a bias corrected estimator is useful for comparison purposes and may offer appreciable improvements in mean-squared error relative to the uncorrected MLE.

Approaches to bias correction may broadly be considered as falling into one of two general categories: simulation approaches and analytic approaches. Simulation approaches, such as the jackknife and bootstrap, rely on resampling and recomputation to estimate the bias. Analytic approaches rely on deriving an analytic expression for the bias and then directly estimating the components of this expression; see, for example, Bartlett (1953) in the scalar

*Date*: This draft: 22 Dec 2005. First draft: 26 Oct 2005.

†University of Chicago, Graduate School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637. Email: chansen1@gsb.uchicago.edu.

case and Shenton and Wallington (1962) and Cox and Snell (1968) in the multivariate case. Specifically, analytic corrections proceed from an expansion of the MLE as

$$\widehat{\theta}_{ml} = \theta_0 + \frac{\psi}{\sqrt{T}} + \frac{b(\theta)}{T} + o_p\left(T^{-1}\right),$$

where $\psi$ is a mean zero random variable satisfying an appropriate central limit theorem. When an analytic solution for $b(\theta)$ is available, a natural bias corrected estimator is $\widehat{\theta}_{bc} = \widehat{\theta}_{ml} - \frac{b(\widehat{\theta}_{ml})}{T}$. An alternative, proposed by Firth (1993), is to obtain a similar bias correction for the scores and define $\widehat{\theta}_{BC}$ as the solution to a bias-corrected score function. In time series and panel settings, bias corrections have been developed for autoregressive models by, for example, Shaman and Stine (1988) and for panel data models with fixed effects by Hahn and Newey (2004) and Hahn and Kuersteiner (2004). In both cases, first order bias correction offers potentially substantial improvement in terms of MSE and inference properties.

Because they are based on an underlying likelihood model, Bayes estimators will also display this frequentist bias. While frequentist bias is unlikely to be of great concern to Bayesian practitioners, there are interesting relationships between frequentist bias-corrections and certain Bayesian priors. For example, Firth (1993) makes the observation that for regular exponential likelihoods, his bias correction of the scores is equivalent to penalizing the log likelihood by one half times the log determinant of the Fisher information matrix, which is better known in the Bayesian literature as the Jeffreys (1946) invariant prior. We argue below that thinking about bias-reduction may offer a potentially useful approach to formulating Bayesian priors.

In part due to the controversy and practical difficulties with formulating prior distributions, the Bayesian literature has explored structural rules for determining priors; see Kass and Wasserman (1996) for an excellent survey of this literature. The resulting priors, often referred to as objective Bayesian priors, formalize the idea of a non-informative prior and are often constructed to endow the resulting estimation procedure with desirable properties. For example, the Jeffreys prior makes the resulting posterior probabilities invariant to reparameterization of the likelihood model. Another example is the class of matching priors proposed by Welch and Peers (1963) and further explored by Ghosh and Mukerjee (1992) and Mukerjee and Dey (1993) among others, which match Bayesian posterior probabilities with frequentist coverage probabilities to higher order asymptotically.

In this paper, we propose a class of likelihood corrections which remove the first-order frequentist bias for two important likelihood estimators: the posterior mode[1] and the posterior mean. We provide general asymptotic expansions for the biases in these two estimators and specify regularity conditions under which our corrections remove first order bias asymptotically. Viewed as a frequentist bias correction, our approach is similar to Firth (1993) but works by correcting the likelihood rather than the scores.

Our likelihood corrections may also be interpreted as objective Bayesian priors. Similar to the development of matching priors in Mukerjee and Dey (1993) and other objective Bayesian priors, our likelihood correction is defined implicitly as the solution to a set of differential equations involving the scores and their derivatives. As these equations will not always be solvable, we also provide a simple and easily implementable data-dependent approximation for the likelihood correction and show that it also removes first order bias asymptotically. This approximation involves only the sample information matrix and outer product of scores, which are already widely used by practitioners conducting frequentist inference.

The next section provides a brief overview of the likelihood models and estimators we consider. In Section 3, we give a rigorous development of asymptotic expansions for the first order bias in the posterior mode and the posterior mean. In Section 4, we construct objective Bayesian priors based on these expansions and prove that the posterior mean and mode under these priors are higher order unbiased. The particular case of panel data models with fixed effects is discussed in Section 4.3. We illustrate our approach with two example models in Section 5 and present a brief Monte Carlo study in Section 6. Section 7 concludes.

## 2. Model and Estimators

Suppose $\{(y_t, x_t) : t = 1, 2, ...\}$ is a sequence of random vectors with $x_t \in \mathcal{X}_t$ and $Y_t \in \mathcal{Y}_t$, and let $w_t = (y_t, x_t)$. Suppose that the conditional density of $y_t$ given $x_t$ with respect to a measure $v_t(dy)$ is given by $f_t(y_t|x_t, \theta_0)$ where $\theta \in \Theta$ is a $p \times 1$ vector of parameters and $\theta_0$ denotes the value of $\theta$ that corresponds to the actual data generating process. We assume that the model is dynamically complete in the sense that $f_t(y_t|x_t, \Phi_{t-1}, \theta_0) = f_t(y_t|x_t, \theta_0)$

[1]Under a flat prior, the posterior mode is the MLE.

where $\Phi_{t-1}$ is the $\sigma$-field representing the past history of the system and note that this definition implicitly allows $x_t$ and $\Phi_{t-1}$ to overlap.

Let $\ell_t(\theta) = \ell_t(w_t, \theta) \equiv \log f_t(y_t|x_t, \theta)$ be the conditional log-likelihood for the $t^{th}$ observation, and let $L_T(\theta) = \prod_{t=1}^{T} \exp\{\ell_t(\theta)\}$ be the likelihood of the sample. Then given a possibly data-dependent prior $\pi_T(\theta) = \pi_T(\theta, w_1, ..., w_T)$ for $\theta$, the posterior distribution for $\theta$ is

$$p_T(\theta|w) = \frac{L_T(\theta)\pi_T(\theta)}{\int_\Theta L_T(\theta)\pi_T(\theta)d\theta}.$$

We note that this formulation of the prior clearly incorporates the usual case of a fixed prior that does not depend on the data where $\pi_T(\theta, w_1, ..., w_T) = \pi(\theta)$ and also incorporates cases where the prior does not depend on the data but does depend on $T$ as in the Jeffreys prior for the AR(1) model where the absolute value of the autoregressive coefficient may be greater than one as derived by Phillips (1991).

In this paper, we focus on two potential Bayesian estimators that also correspond to measures of the location of the posterior distribution: the joint posterior mode and the posterior mean. The joint posterior mode is given by

$$\widehat{\theta}^M = \arg\max_\theta \, p_T(\theta|w) = \arg\max_\theta \, \ell_T(\theta) + \gamma_T(\theta) \tag{2.1}$$

where $\ell_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta) = \log[L_T(\theta)]$ and $\gamma_T(\theta) = \log[\pi_T(\theta)]$, and the posterior mean is defined as

$$\widehat{\theta}^E = E_{p(\theta|w)}[\theta] = \int_\Theta \theta p(\theta|w)d\theta \tag{2.2}$$

when this expectation exists.

We focus on the posterior mean and mode for a number of reasons. Both estimators provide sensible measures of the location of the posterior distribution and have been used extensively in both Bayesian and classical settings. In Bayesian estimation, the posterior mode and mean correspond to the use of 0-1 and quadratic loss functions, respectively. In a classical setting, the posterior mode may be interpreted as a penalized maximum likelihood estimator with the prior serving as a penalty function. The posterior mean and mode are also analytically tractable and are naturally examined using the higher order asymptotics presented below.

## 3. Asymptotic Expansions

We find the asymptotic bias for $\widehat{\theta}^M$ and $\widehat{\theta}^E$ using stochastic expansions for each estimator. We use these expansions in the next section to define our bias reducing priors. However, we argue below that the expansions may be useful in other settings; for example, in comparing the biases induced by different types of priors.

These expansions allow us to express each estimator as

$$\widehat{\theta}^j = \theta_0 + \psi_1/\sqrt{T} + \psi_2^j/T + R_T$$

for $j \in \{M,\ E\}$ where $\psi_1$ is a mean zero random variable that follows a central limit theorem, $\psi_2^j$ is a random variable that may not have zero mean, and $R_T = o_p(1/T)$. We can then define an "approximate" estimator $\widetilde{\theta}^j = \theta_0 + \psi_1/\sqrt{T} + \psi_2^j/T$ with expectation $\mathrm{E}[\widetilde{\theta}^j] = \theta_0 + \mathrm{E}[\psi_2^j]/T$. In this sense, we can understand $\mathrm{E}[\psi_2]/T$ as the higher-order or approximate bias in the estimator $\widehat{\theta}^j$. Throughout the remainder of the paper, we are referring to this form of higher-order bias whenever we discuss the bias of an estimator.

Before turning to a heuristic derivation of the higher-order expansions, it will be useful to define some notation. Let superscripts denote differentiation such that $\ell_t^\theta(\theta) = \frac{\partial \ell_t(\theta)}{\partial \theta}$, $\ell_t^{\theta\theta}(\theta) = \frac{\partial^2 \ell_t(\theta)}{\partial \theta \partial \theta'}$, $\ell_t^{\theta\theta_j}(\theta) = \frac{\partial^3 \ell_t(\theta)}{\partial \theta \partial \theta' \partial \theta_j}$, etc. Define $S_T(\theta) = \frac{1}{T}\sum_{t=1}^T \ell_t^\theta(\theta)$, $\widehat{H}_T(\theta) = \frac{1}{T}\sum_{t=1}^T \ell_t^{\theta\theta}(\theta)$, and $\widehat{H}_T^{\theta_j}(\theta) = \frac{1}{T}\sum_{t=1}^T \ell_t^{\theta\theta\theta_j}(\theta)$. Similarly, define $H_T(\theta) = \frac{1}{T}\sum_{t=1}^T \mathrm{E}[\ell_t^{\theta\theta}(\theta)]$ and $H_T^{\theta_j}(\theta) = \frac{1}{T}\sum_{t=1}^T \mathrm{E}[\ell_t^{\theta\theta\theta_j}(\theta)]$. Finally, note that arguments are suppressed when the functions are evaluated at $\theta_0$; e.g. $S_T = S_T(\theta_0) = \frac{1}{T}\sum_{t=1}^T \ell_t^\theta(\theta_0)$.

The expansion for the posterior mode may be obtained in a straightforward fashion from the first-order conditions of the optimization problem (2.1). Specifically, $\widehat{\theta}^M$ satisfies

$$0 = \frac{1}{T}\frac{\partial}{\partial \theta}\left(\ell_T(\widehat{\theta}^M) + \gamma_T(\widehat{\theta}^M)\right) = S_T(\widehat{\theta}^M) + \frac{1}{T}\gamma_T^\theta(\widehat{\theta}^M). \tag{3.1}$$

Expanding the left hand side about $\widehat{\theta} = \theta_0$ produces

$$\begin{aligned} S_T(\widehat{\theta}^M) + \frac{1}{T}\gamma_T^\theta(\widehat{\theta}^M) = S_T + H_T(\widehat{\theta}^M - \theta_0) + (\widehat{H}_T - H_T)(\widehat{\theta}^M - \theta_0) + \frac{1}{T}\gamma_T^\theta \\ + \frac{1}{2}\sum_{j=1}^p \widehat{H}_T^{\theta_j}(\widehat{\theta}^M - \theta_0)(\widehat{\theta}_j^M - \theta_{j0}) + o_p(T^{-1}) \end{aligned} \tag{3.2}$$

where $\theta_j$ is the $j^{th}$ element of $\theta$.

Using (3.1) and (3.2), subtracting $H_T(\widehat{\theta}^M - \theta_0)$ from both sides of the equation, and multiplying by $-(H_T)^{-1}$ yields

$$\widehat{\theta}^M - \theta_0 = -(H_T)^{-1}S_T - (H_T)^{-1}(\widehat{H}_T - H_T)(\widehat{\theta}^M - \theta_0) - (H_T)^{-1}\frac{1}{T}\gamma_T^{\theta} \tag{3.3}$$
$$- \frac{1}{2}\sum_{j=1}^{p}(H_T)^{-1}\widehat{H}_T^{\theta_j}(\widehat{\theta}^M - \theta_0)(\widehat{\theta}_j^M - \theta_{j0}) + o_p(T^{-1}).$$

Under regularity conditions, all of the terms in (3.3) are $o_p(T^{-1/2})$ except for the leading term, $-(H_T)^{-1}S_T$, which is $O_p(T^{-1/2})$ and follows a central limit theorem when multiplied by $T^{1/2}$. The usual first-order expansion,

$$\widehat{\theta}^M - \theta_0 = -(H_T)^{-1}S_T + o_p(T^{-1/2}) = \frac{1}{\sqrt{T}}\psi_1 + o_p(T^{-1/2}) \tag{3.4}$$

where $\psi_1 \xrightarrow{d} N(0, -(\lim_{T\to\infty} H_T)^{-1})$, then follows. Finally, replacing the $\widehat{\theta}^M - \theta_0$ terms on the right-hand side of (3.3) with the expression in (3.4) gives the expansion for the posterior mode:

$$\widehat{\theta}^M - \theta_0 = T^{-1/2}\psi_1 - T^{-1}(H_T)^{-1}\bar{\gamma}_T^{\theta} - T^{-1}(H_T)^{-1}\left[\sqrt{T}\left(\widehat{H}_T - H_T\right)\right]\psi_1 \tag{3.5}$$
$$- \frac{1}{2}T^{-1}\sum_{j=1}^{k}(H_T)^{-1}H_T^{\theta_j}\psi_1\psi_{j1} + o_p(T^{-1})$$
$$= T^{-1/2}\psi_1 + T^{-1}\psi_2^M + o_p(T^{-1})$$

where $\psi_{j1}$ is the $j^{th}$ element of $\psi_1$ and $\bar{\gamma}_T^{\theta}$ is a nonstochastic function with $\bar{\gamma}_T^{\theta} - \gamma_T^{\theta} = o_p(1)$. It follows that the higher order bias of $\widehat{\theta}^M$ is $\mathrm{E}[\psi_2^M]/T$.

It is also straightforward to derive a similar expansion for the posterior mean, $\widehat{\theta}^E$. Using the Laplace expansion of the posterior mean obtained in Kass, Tierney, and Kadane (1990), we have

$$\widehat{\theta}^E = \widehat{\theta}^M + \frac{1}{2T}\left(\widehat{H}_T(\widehat{\theta}^M)\right)^{-1}\sum_{j=1}^{p}\mathrm{trace}\left[\widehat{H}_T^{\theta_j}(\widehat{\theta}^M)\left(\widehat{H}_T(\widehat{\theta}^M)\right)^{-1}\right]e_j + o_p(T^{-1}), \tag{3.6}$$

where $e_j$ is the $j^{th}$ unit vector. It may then be shown that (3.6) may be written as

$$\widehat{\theta}^E - \theta_0 = \widehat{\theta}^M - \theta_0 + \frac{1}{2T}(H_T)^{-1}\sum_{j=1}^{p}\mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j + o_p(T^{-1})$$

after further expanding about $\widehat{\theta}^M = \theta_0$. Finally, replacing $\widehat{\theta}^M - \theta_0$ with the expansion in (3.5) gives

$$\widehat{\theta}^E - \theta_0 = T^{-1/2}\psi_1 + T^{-1}\psi_2^E + o_p(T^{-1}) \tag{3.7}$$

where

$$\begin{aligned}\psi_2^E = &-T^{-1}(H_T)^{-1}\bar{\gamma}_T^\theta - T^{-1}(H_T)^{-1}\left[\sqrt{T}\left(\widehat{H}_T - H_T\right)\right]\psi_1 \\ &- \frac{1}{2}T^{-1}\sum_{j=1}^{k}(H_T)^{-1}H_T^{\theta_j}\psi_1\psi_{j1} + \frac{1}{2T}(H_T)^{-1}\sum_{j=1}^{p}\text{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j\end{aligned} \tag{3.8}$$

and $\mathrm{E}[\psi_2^E]/T$ is the $O(1/T)$ bias in the posterior mean.

In order to insure that the expansions given in (3.5) and (3.7) are valid, we impose the following set of conditions. In the following, let $\Delta^j$ represent a vector of all distinct partial derivatives with respect to $\theta$ of order $j$. Also, throughout the remainder of the paper, we let $\|A\| = \text{trace}(A'A)^{1/2}$ denote the usual Euclidean norm of a matrix $A$.

**Assumption 1.** *For the model defined in Section 2, suppose that (i) $\theta_0 \in int(\Theta)$ where $\Theta$ is a compact, separable metric space; (ii) $\{w_t,\ t = 1, 2, ...\}$ is a mixing sequence that satisfies, for $A_t = \sigma(w_t, w_{t-1}, ...)$, $B_t = \sigma(w_t, w_{t+1}, ...)$, and $\alpha(m) = \sup_t \sup_{\{A\in A_t,\ B\in B_t\}} |P(A \cap B) - P(A)P(B)|$, $\alpha(m) = O(m^{\frac{-r}{r-2}-\epsilon})$ for some $\epsilon > 0$ and $r > 2$; (iii) there exists a function $M_t(w_t)$ such that for $0 \le j \le 6$, all $w_t \in \mathcal{W}_T = \mathcal{Y}_t \times \mathcal{X}_t$, and all $\theta \in \mathcal{G}$ where $\mathcal{G}$ is an open, convex set containing $\Theta$, $\Delta^j\ell_t(\theta)$ exists, $\sup_{\theta\in\mathcal{G}}\|\Delta^j\ell_t(\theta)\| \le M_t(w_t)$, and $\sup_t \mathrm{E}\|M_t(w_t)\|^{r+\delta} \le M < \infty$ for some $\delta > 0$; (iv) $\theta_0$ is the unique maximizer of $\bar{Q}(\theta) = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathrm{E}[\ell_t(\theta)]$; (v) $H_T$ has maximum eigenvalue $\lambda_{\max} \le \delta < 0$ and minimum eigenvalue $\lambda_{\min} \ge -M > -\infty$ uniformly in $T$; (vi) $H_0 \equiv \lim_{T\to\infty} H_T$ exists and is nonsingular, and $H_0^{\theta_j} \equiv \lim_{T\to\infty} H_T^{\theta_j}$ exists; (vii) $\exp\{\gamma_T(\theta)\} > 0$ for all $\theta \in \Theta$; (viii) for $0 \le j \le 6$ $\sup_{\theta\in\mathcal{G}}\mathrm{E}\|\Delta^j\gamma_T(\theta)\| = o(T^{1/2})$; (ix) there is a nonstochastic function $\bar{\gamma}_T^\theta$ that satisfies $\gamma_T^\theta - \bar{\gamma}_T^\theta = o_p(1)$.*

Conditions (i)-(vi) of Assumption 1 are quite similar to other assumptions made to justify higher order expansions; see, for example, Rilstone, Srivastava, and Ullah (1996) and Newey and Smith (2001). The chief difference is the imposition of somewhat stronger smoothness conditions which are used in establishing the validity of the Laplace expansion used to derive (3.5). Assumption 1 also allows for dependent and potentially nonstationary data instead

of imposing that the data are iid. The existence of the limits in (vi) could be relaxed at the cost of more complicated proof and notation.

Conditions (vii)-(ix) guarantee that the prior is well-behaved. Condition (vii) requires that the prior is supported over the entire parameter space, and Condition (viii) imposes sufficient smoothness on the prior to verify the higher order expansions. In addition, the moment bound in Condition (viii) ensures that the likelihood dominates the prior asymptotically. Conditions (vii)-(ix) are trivially satisfied by most common fixed priors employed in practice.

The conditions imposed in Assumption 1 are sufficient to allow us to state the following results.

**Theorem 1.** *If Assumption 1 is satisfied,* $\widehat{\theta}^M = \theta_0 + \psi_1/T^{1/2} + \psi_2^M/T + o_p(1/T)$ *where* $\widehat{\theta}^M$ *is the joint posterior mode for* $\theta$ *and* $\psi_1$ *and* $\psi_2^M$ *are defined in (3.4) and (3.5). If, in addition, there is some nonnegative integer* $T_0$ *such that the posterior based on a sample of* $T_0$ *observations* $w^{T_0} = \{w_1, ..., w_{T_0}\}$ $p_{T_0}(\theta|w^{T_0})$ *exists and has finite expectation,* $\widehat{\theta}^E = \theta_0 + \psi_1/T^{1/2} + \psi_2^E/T + o_p(1/T)$ *where* $\widehat{\theta}^E$ *is the posterior mean for* $\theta$ *and* $\psi_1$ *and* $\psi_2^E$ *are defined in (3.4) and (3.8).*

Theorem 1 simply formalizes the previous heuristic argument by verifying that the derived expansions are valid under a set of regularity conditions. The higher-order biases of the estimators then follow immediately by taking expectations of the second term in each expansion and are presented in the following result for convenience.

**Corollary 1.1.** *Under Assumption 1, the higher-order bias of the posterior mode,* $\widehat{\theta}^M$*, is*

$$\mathrm{E}[\psi_2^M]/T = -T^{-1}(H_T)^{-1}\bar{\gamma}_T^{\theta} - T^{-1}(H_T)^{-1}\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \frac{1}{2}T^{-1}\sum_{j=1}^{k}(H_T)^{-1}\mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j. \tag{3.9}$$

*If, in addition, there is some nonnegative integer* $T_0$ *such that the posterior based on a sample of* $T_0$ *observations* $w^{T_0} = \{w_1, ..., w_{T_0}\}$ $p_{T_0}(\theta|w^{T_0})$ *exists and has finite expectation,*

*the higher-order bias of the posterior mean, $\widehat{\theta}^E$, is*

$$\mathrm{E}[\psi_2^E]/T = -T^{-1}(H_T)^{-1}\bar{\gamma}_T^{\theta} - T^{-1}(H_T)^{-1}\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + T^{-1}\sum_{j=1}^{k}(H_T)^{-1}\mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j. \tag{3.10}$$

The higher-order bias formulae may be of interest for a number of reasons. For any Bayesian prior satisfying the regularity conditions in Assumption 1, the expressions can be used to determine the approximate bias in the joint posterior mode or posterior mean. In situations where there is little available prior information or the available prior information may be reasonably represented in more than one way, comparisons of the resulting biases may be useful for comparing the different priors. It is also worth noting that for many choices of prior, the resulting bias to the posterior mode will be smaller than the bias of the maximum likelihood estimator which corresponds to the posterior mode under a flat prior.

From the expressions for the bias of the posterior mode and mean given in (3.9) and (3.10), we can also see that there are three sources of higher order bias: bias resulting from correlation between the influence function and approximate second derivate matrix, $\widehat{H}_T$; bias resulting from curvature in the model captured by the last term in each expression; and bias from the prior. While the first two sources of bias are determined by the likelihood, the third is under the control of the researcher. In particular, in some settings where there is little or no available prior information, a researcher may wish to define a prior that reduces the asymptotic bias. (3.9) and (3.10) then provide a way to define objective Bayesian priors, where the objective is to reduce the higher-order frequentist bias of the resulting posterior mean or joint posterior mode.

## 4. Bias-Reducing Priors

In the previous section, we presented stochastic expansions for the joint posterior mode and posterior means of the parameters in a likelihood model and obtained the $O(1/T)$ bias of the posterior mode and posterior mean. In both cases, the bias depends on the form of the prior. In this section, we show that this dependence implicitly defines a class of bias reducing priors. Our bias-reducing priors recenter the resulting posterior distribution such that the $O(1/T)$ bias is removed from either the posterior mode or the posterior mean.

We consider two approaches to constructing bias reducing priors. In the first approach, certain components of the bias are computed analytically and equated with the derivatives of the prior, resulting in a system of differential equations. When the differential equations may be solved for each possible value of $\theta$ and the solution does not depend on the data, we call the solution an "exact" bias-reducing prior. In practice, such a solution may not be available in closed form or may depend on the data. We therefore provide a simple and tractable data dependent prior that solves the differential equations asymptotically and hence also removes the first order bias from the posterior mean or mode.

4.1. **Closed Form Bias-Reducing Priors.** Using the stochastic expansions and resulting higher-order bias formulae presented in the preceding section, we may define objective Bayesian priors that remove the higher-order bias from the mode or mean of the resulting posterior. As with many objective Bayesian priors, these "bias-reducing" priors may be defined as the solution to a set of differential equations; see Kass and Wasserman (1996) for an overview of other objective Bayesian approaches.

Specifically, we can set expression (3.9) equal to 0 and solve for $\bar{\gamma}_T^\theta$ to obtain

$$\bar{\gamma}_T^\theta = -\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \frac{1}{2}\sum_{j=1}^{k} \mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right] e_j. \tag{4.1}$$

Then, recalling that $\gamma_T^\theta - \bar{\gamma}_T^\theta = o_p(1)$ and $\gamma_T^\theta = \frac{\partial}{\partial\theta}\log[\pi_T(\theta)]$, we see that any prior (potentially data-dependent) whose derivative converges in probability to the expression on the left-hand side of (4.1) will remove the higher-order bias from the posterior mode. Similarly using (3.10), we have that any prior whose derivative converges in probability to $\bar{\gamma}_T^\theta$ where

$$\bar{\gamma}_T^\theta = -\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \sum_{j=1}^{k} \mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right] e_j \tag{4.2}$$

will remove the $O(1/T)$ bias from the posterior mean.

Equations (4.1) and (4.2) provide sets of differential equations that define priors that will asymptotically remove the higher-order bias from the posterior mode and mean respectively. One approach to obtaining such priors is to compute the expectations on the right-hand side of equations (4.1) and (4.2) analytically; note that after these expectations are taken the right hand sides of both equations depend only on $\theta$ and, in some cases, the sample size

$T$. Solutions to the resulting set of differential equations, if they exist, may be used as bias reducing priors.

**Theorem 2.** *Suppose Assumption 1 (i)-(v) are satisfied and there exists a prior $\pi_T(\theta)$ such that $\pi_T(\theta)$ is six times continuously differentiable with all partial derivatives bounded and $\pi_T(\theta) > 0$ for all $\theta \in \Theta$. If $\pi_T(\theta)$ satisfies*

$$\frac{\partial}{\partial \theta} \log[\pi_T(\theta)] = -\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \frac{1}{2}\sum_{j=1}^{k} \mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right] e_j,$$

*then the resulting joint posterior mode is higher-order unbiased; and if $\pi_T(\theta)$ satisfies*

$$\frac{\partial}{\partial \theta} \log[\pi_T(\theta)] = -\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \sum_{j=1}^{k} \mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right] e_j,$$

*then the resulting posterior mean is higher-order unbiased.*

In many cases, notably when data are iid, the right-hand side of (4.1) and (4.2) will not depend on $T$. In these cases, the differential equations can be used to define fixed priors $\pi(\theta)$. Otherwise, the priors will generally depend on the sample size and $\theta$ but will not otherwise depend on the data.

While this approach to deriving a bias-reducing prior defines a "true" prior in the sense that the prior does not depend on the data, it may difficult to implement for a number of reasons. Computing the expectations on the right hand sides of (4.1) and (4.2) analytically may be quite difficult, and except in simple models, the expressions may not be available in closed-form. Also, even with the expectations computed, the resulting differential equations may not have closed form solutions. The resulting prior would then have to be evaluated numerically, further complicating implementation. It is also important to note that the systems of differential equations defined in (4.1) and (4.2) are overdetermined when $\dim(\theta) > 1$ and may be inconsistent.[2] In the next section, we consider an alternative approach to

[2]In these cases, if there is a single parameter of interest, one may consider only the row of the system corresponding to that parameter. Focusing on one parameter will yield a single ordinary differential equation which will always be soluble and whose solution will often satisfy the regularity conditions stated in Theorem 2. Alternatively, if one is interested in the entire parameter vector, the data-dependent prior defined below may be employed.

constructing bias correcting priors. Although the resulting priors will in general be data dependent, they are always available and are simple to compute.

4.2. **Data Dependent Bias-Reducing Priors.** The condition employed in the previous section to define the prior, namely that the derivative of the prior exactly satisfy the differential equations (4.1) and (4.2) for all values of $\theta$ in the parameter space, is much stronger than is actually needed to define a bias-reducing prior. In particular, all that is necessary is that the derivative of the log of the prior approximately satisfy the differential equation in the sense that the derivative of the prior converges in probability to the expression on the right-hand side of the differential equations when evaluated at the true parameter value $\theta_0$. In this section, we consider one such approximate solution that is particularly convenient as it is always available and is simple to compute.

The "approximate" prior that we consider is data-dependent; that is, it depends on $\{w_1, ..., w_T\}$.[3] In situations where a prior satisfying the conditions of Theorem 2 exists for the mode or mean, the data-dependent prior may be viewed as an approximation to the prior defined in Theorem 2 in that it also satisfies the differentiability and boundedness conditions of Theorem 2 and satisfies the differential equation when evaluated at $\theta_0$. In situations where the differential equations are inconsistent, it may be viewed as an approximation to a prior that removes the bias from a single parameter[4] and satisfies the conditions of Theorem 2 in the sense discussed above that has the additional appealing property of removing the higher-order bias from the remaining elements of the parameter vector as well.

Specifically, as an approximate solution to (4.1) when the likelihood is globally concave,[5] we consider

$$\pi_T^M(\theta) = \exp\{\frac{1}{2}\text{tr}\left[\widehat{H}^{-1}(\theta)\widehat{V}(\theta)\right]\} \tag{4.3}$$

[3]The use of data-dependent priors in objective Bayes approaches is not new; see, for example, Wasserman (2000), Reid, Mukerjee, and Fraser (2002), and Sweeting (2004).

[4]As discussed above, such a prior may be defined by considering the appropriate row of the differential equation (4.1) or (4.2).

[5]The general case is somewhat more complicated and is detailed in Appendix A.

where

$$\widehat{V}(\theta) = \frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^{\theta}(\theta)(\ell_{t-j}^{\theta}(\theta))', \tag{4.4}$$

$m \to \infty$ is a bandwidth parameter that satisfies $m/T^{1/2} \to 0$, and $K(\cdot)$ is a kernel function with $K(0) = 1$ and $\int_{-\infty}^{\infty} |K(x)|dx < \infty$. That (4.3) is an approximation to a prior satisfying (4.1) in the sense that the derivative of the logarithm of (4.3) satisfies (4.1) follows by differentiating (4.3), evaluating at $\theta = \theta_0$, and taking limits as $T \to \infty$. Similarly, as an approximate solution to (4.2) when the likelihood is globally concave,[6], we have

$$\pi_T^E(\theta) = |-\widehat{H}|^{1/2}\exp\{\frac{1}{2}\text{tr}\left[\widehat{H}^{-1}(\theta)\widehat{V}(\theta)\right]\} \tag{4.5}$$

where $\widehat{V}(\theta)$ is defined as in (4.4). As above, one may verify that $\pi_T^E(\theta)$ satisfies (4.2) by differentiating, evaluating at $\theta_0$, and taking limits as $T \to \infty$.

As noted above, the approximate priors are appealing because they exist and satisfy the derivative condition given by (4.1) or (4.2) quite generally. In addition, they involve only quantities that are used in forming the usual frequentist asymptotic standard errors: the approximate scores and hessian of the maximum likelihood problem. These quantities are readily available by simply differentiating the maximum likelihood objective function. As such, calculating the approximate priors is quite straightforward.

To formally verify that $\pi_T^M(\theta)$ and $\pi_T^E(\theta)$ respectively remove the bias from the joint posterior mode and posterior mean, we impose the following additional regularity conditions.

**Assumption 2.** *For the model defined in Section 2, suppose that (i) for all $\theta \in \Theta$, $\widehat{H}_T$ has maximum eigenvalue $\widehat{\lambda}_{\max} \leq \delta < 0$ and minimum eigenvalue $\widehat{\lambda}_{\min} \geq -M > -\infty$; (ii) $\widehat{V}_T(\theta)$ is positive semi-definite for all $\theta \in \Theta$; (iii) $\{w_t,\ t = 1, 2, ...\}$ is a mixing sequence that satisfies $\alpha(m) = O(m^{\frac{-3r}{r-2}-\epsilon})$ for some $\epsilon > 0$ and $r > 2$; and (iv) there exists a function $M_t(w_t)$ such that for $0 \leq j \leq 8$, all $w_t \in \mathcal{W}_T = \mathcal{Y}_t \times \mathcal{X}_t$, and all $\theta \in \mathcal{G}$ where $\mathcal{G}$ is an open, convex set containing $\Theta$, $\Delta^j \ell_t(\theta)$ exists, $\sup_{\theta\in\mathcal{G}} \|\Delta^j \ell_t(\theta)\| \leq M_t(w_t)$, and $\sup_t \text{E}\|M_t(w_t)\|^{4r+\delta} \leq M < \infty$ for some $\delta > 0$.*

Conditions (i)-(iv) of Assumption 2 are sufficient to guarantee that Conditions (vii)-(ix) of Assumption 1 are satisfied. Assumption 2.(i) imposes concavity on the model over the

[6]The general case is covered in Appendix A.

parameter space; this condition is relaxed in Appendix A where we present approximate priors that will remove the $O(1/T)$ bias from the posterior mode or mean when the likelihood is not globally concave. Condition (ii) requires that the estimate of the second moment of the scores be positive semi-definite for all $\theta$. Appropriate choice of the kernel function used in defining $\widehat{V}_T(\theta)$ will guarantee that this condition is satisfied; see, for example, White (2001) Lemma 6.22.[7] Conditions (iii) and (iv) strengthen the mixing and moment conditions sufficiently to guarantee that Assumption 1.(vii) is satisfied and that the estimator $\widehat{V}$ at $\theta_0$ converges in probability to $V$.

Under the conditions of Assumption 2, we have that $\pi_T^M(\theta)$ and $\pi_T^E(\theta)$ remove the higher-order bias from the joint posterior mode and posterior mean.

**Theorem 3.** *If $T \to \infty$ and $m \to \infty$ such that $m/T^{1/2} \to 0$ and Assumption 2 holds, Conditions (vii)-(ix) of Assumption 1 are satisfied by $\log[\pi_T^M(\theta)]$ and $\log[\pi_T^E(\theta)]$ and*

$$\frac{\partial}{\partial\theta}\log[\pi_T^M(\theta)] - \left(-\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \frac{1}{2}\sum_{j=1}^{k}\mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j\right) = o_p(1)$$

*and*

$$\frac{\partial}{\partial\theta}\log[\pi_T^E(\theta)] - \left(-\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \sum_{j=1}^{k}\mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j\right) = o_p(1).$$

*It follows that $\pi_T^M(\theta)$ removes the higher-order bias from the posterior mode and that $\pi_T^E(\theta)$ removes the higher-order bias from the posterior mean.*

4.3. **Simplifications for Panel Data with Fixed Effects.** Bias reduction is clearly most relevant in cases where estimators have a large amount of bias. One case where this seems especially prevalent is in models where the number of parameters goes to infinity with the sample size in such a way that the amount of data per parameter does not accumulate, that is in models with incidental parameters.[8] In econometrics, perhaps the most prevalent case where this appears is in panel data models with fixed individual specific effects. Recently, there has also been a great deal of work in reducing bias of estimators of common parameters in panel data models with fixed individual specific effects; see, for example, the review article by Arellano and Hahn (2005) which contains an excellent overview of recent developments.

[7]Newey and West (1987) and Andrews (1991) provide additional discussion and examples.

[8]See Neyman and Scott (1948) for an early use of this terminology.

For concreteness,[9] suppose we are interested in a panel data model where $i = 1, ..., N$ indexes individuals and $t = 1, ..., T$ indexes time periods and the conditional density of $y_{it}$ given $x_{it}$ and individual specific effects $\alpha_{i0}$ is given by $f(y_{it}|x_{it}, \alpha_{i0}, \theta_0)$. We suppose that the parameter of interest is $\theta_0$ and the $\alpha_{i0}$ are unobserved nuisance parameters that may be arbitrarily correlated to the $T \times 1$ vector of covariates for individual $i$, $x_i$.

The focus on bias in models with incidental parameters stems from the observation that the bias in estimators of the common parameters, $\theta$, generally diminishes at a rate equal to $1/T$ while the variance of estimators of $\theta$ diminishes at a rate equal to $1/NT$.[10] This means, particularly in applications where $N$ is large and $T$ is small, that bias is typically the dominant factor in mean squared error and suggests that estimation and inference properties may be substantially improved by accounting for this bias.

The approach to bias-reduction via priors outlined in the previous sections may be immediately adapted to the present setting by considering the relevant sample size to be $T$, the number of observations used to estimate the $\alpha_i$. In this case, the priors defined in Sections 4.1 and 4.2 will remove the $O(1/T)$ bias from the estimators. Removing the bias to this order will generally result in estimators that are consistent and have correctly centered asymptotic distributions in asymptotics where $N \to \infty$, $T \to \infty$, and $N/T \to \rho$ for some constant $\rho$; see Hahn and Kuersteiner (2004). In finite samples, this correction may substantially improve inference properties of the resulting estimators.

A drawback of using the full priors defined in the previous sections in this setting is that the high dimensional parameter space may complicate computation of the priors. In addition, conditional independence between $\alpha_i$ and $\alpha_j$ for different individuals $i \neq j$, a general feature of fixed effects panel models that greatly aids computation of Bayesian estimators via Markov Chain Monte Carlo, may be destroyed when the full priors are employed. Fortunately, there are some simplifications which are available in these models when $N$ is relatively large that arise by noting that the $O(1/T)$ bias contains terms that behave like $1/NT$. When $N$ is large, ignoring these terms will not affect the bias-reducing properties of the priors but may greatly simplify their computation and implementation.

[9]The discussion in this section could be readily extended to other models with incidental parameters.

[10]See, for example, Arellano and Hahn (2005), Hahn and Newey (2004), or Hahn and Kuersteiner (2004) for related discussion and formalization of this argument.

Heuristically, we can see how this simplification may be accomplished by considering a simple fixed effects model with $\dim(\alpha_i) = 1$ and stationary data. For stationary models, we have that

$$H_T = H = \begin{pmatrix} N\mathrm{E}[\ell_{it}^{\theta\theta}] & \mathrm{E}[\ell_{it}^{\theta\alpha_i}]\iota_N' \\ \iota_N \mathrm{E}[\ell_{it}^{\alpha_i\theta}] & \mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}] I_N \end{pmatrix}$$

where $\iota_N$ is an $N \times 1$ vector of ones and $I_N$ is the $N \times N$ identity matrix. It follows that

$$\begin{aligned} H^{-1} &= \begin{pmatrix} (N\mathrm{E}[\ell_{it}^{\theta\theta}] - N\frac{\mathrm{E}[\ell_{it}^{\theta\alpha_i}]\mathrm{E}[\ell_{it}^{\alpha_i\theta}]}{\mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}]})^{-1} \equiv H^{11} & -\frac{H^{11}\mathrm{E}[\ell_{it}^{\theta\alpha_i}]\iota_N'}{\mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}]} \\ -\frac{\iota_N \mathrm{E}[\ell_{it}^{\alpha_i\theta}] H^{11}}{\mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}]} & \frac{I_N}{\mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}]} - \frac{\iota_N \mathrm{E}[\ell_{it}^{\alpha_i\theta}] H^{11} \mathrm{E}[\ell_{it}^{\theta\alpha_i}]\iota_N'}{\mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}]^2} \end{pmatrix} \\ &\approx \begin{pmatrix} 0 & 0 \\ 0 & \frac{I_N}{\mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}]} \end{pmatrix}. \end{aligned}$$

Similarly, without imposing stationarity and allowing $\dim(\alpha_i) \geq 1$, we would have

$$H_T^{-1} \approx \begin{pmatrix} 0 & 0 \\ 0 & A_T \end{pmatrix} \tag{4.6}$$

where $A_T$ is a block diagonal matrix with $i^{th}$ block equal to $(H_T^{\alpha_i\alpha_i})^{-1} = (T^{-1}\sum_{t=1}^T \mathrm{E}[\ell_{it}^{\alpha_i\alpha_i}])^{-1}$. We may then plug equation (4.6) into (4.1) and, after some algebra and assuming independence across $i$, obtain

$$\begin{aligned} \bar{\gamma}_T^{\alpha_i} = \mathrm{E}&\left[\frac{1}{T}\sum_{t=1}^T \ell_{it}^{\alpha_i\alpha_i} \sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i}\right] \\ &+ \frac{1}{2}\sum_{j=1}^{\dim(\alpha_i)} \mathrm{E}\left[\widehat{H}_T^{\alpha_i\alpha_i\alpha_{i,j}}\right] \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i} \sum_{t=1}^T ((H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i})'\right] e_j \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} \bar{\gamma}_T^{\theta} = \sum_{i=1}^N \mathrm{E}&\left[\frac{1}{T}\sum_{t=1}^T \ell_{it}^{\theta\alpha_i} \sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i}\right] \\ &+ \sum_{i=1}^N \frac{1}{2}\sum_{j=1}^{\dim(\theta)} \mathrm{E}\left[\widehat{H}_T^{\alpha_i\alpha_i\theta_j}\right] \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i} \sum_{t=1}^T ((H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i})'\right] e_j \end{aligned} \tag{4.8}$$

where $\widehat{H}_T^{\alpha_i\alpha_i\theta_j} = \frac{\partial}{\partial\theta_j} T^{-1}\sum_{t=1}^T \ell_{it}^{\alpha_i\alpha_i}$, $\widehat{H}_T^{\alpha_i\alpha_i\alpha_{i,j}} = \frac{\partial}{\partial\alpha_{i,j}} T^{-1}\sum_{t=1}^T \ell_{it}^{\alpha_i\alpha_i}$, and $\alpha_{i,j}$ is the $j^{th}$ element of $\alpha_i$. Equations (4.7) and (4.8) jointly define a system of differential equations that

defines a bias-reducing prior for the panel case. Note that making use of the simplification, the structure of equations (4.7) and (4.8) imply that any solution will have the form $\log[\pi_T(\alpha_1, ..., \alpha_N, \theta)] = \sum_{i=1}^N \log[\pi_{iT}(\alpha_i, \theta)]$ for some $\pi_{iT}(\alpha_i, \theta)$. That is, conditional on $\theta$, the prior exhibits independence across the $\alpha_i$. This conditional independence, coupled with the conditional independence in the likelihood, greatly simplifies simulation from the posterior via Markov Chain Monte Carlo through blocking schemes (e.g., Gibbs sampling).

If, instead, we want a bias-reducing prior for the posterior mean, we may plug (4.6) into (4.2) to obtain

$$\begin{aligned}\bar{\gamma}_T^{\alpha_i} = \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{it}^{\alpha_i\alpha_i}\sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i}\right] \\ + \frac{1}{2}\sum_{j=1}^{\dim(\alpha_i)}\left(\mathrm{E}\left[\widehat{H}_T^{\alpha_i\alpha_i\alpha_{i,j}}\right]\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i}\sum_{t=1}^T ((H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i})'\right]\right. \\ \left. -\mathrm{trace}[\mathrm{E}\left[\widehat{H}_T^{\alpha_i\alpha_i\alpha_{i,j}}\right](H_T^{\alpha_i\alpha_i})^{-1}]I_{\dim(\alpha_i)}\right)e_j\end{aligned} \tag{4.9}$$

and

$$\begin{aligned}\bar{\gamma}_T^{\theta} = \sum_{i=1}^N \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{it}^{\theta\alpha_i}\sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i}\right] \\ + \sum_{i=1}^N\frac{1}{2}\sum_{j=1}^{\dim(\theta)}\left(\mathrm{E}\left[\widehat{H}_T^{\alpha_i\alpha_i\theta_j}\right]\mathrm{E}\left[\frac{1}{T}\sum_{t=1}^T (H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i}\sum_{t=1}^T ((H_T^{\alpha_i\alpha_i})^{-1}\ell_{it}^{\alpha_i})'\right]\right. \\ \left. -\mathrm{trace}[\mathrm{E}\left[\widehat{H}_T^{\alpha_i\alpha_i\theta_j}\right](H_T^{\alpha_i\alpha_i})^{-1}]I_{\dim(\theta)}\right)e_j.\end{aligned} \tag{4.10}$$

As with the solution for the posterior mean, any solution to this set of differential equations will be of the form $\log[\pi_T(\alpha_1, ..., \alpha_N, \theta)] = \sum_{i=1}^N \log[\pi_{iT}(\alpha_i, \theta)]$.

As in the general case, simple data-dependent approximations to the solutions of these differential equations are also available. In particular, we have that

$$\pi_{iT}^M(\theta, \alpha_i) = \exp\{\frac{1}{2}\mathrm{trace}[(\widehat{H}_T^{\alpha_i\alpha_i}(\theta, \alpha_i))^{-1}\widehat{V}^{\alpha_i}(\theta, \alpha_i)]\} \tag{4.11}$$

satisfies (4.7) and (4.8) for $\pi_{iT}^M(\theta, \alpha_i)$ defined above, $\widehat{H}_T^{\alpha_i\alpha_i}(\theta, \alpha_i) = \frac{1}{T}\sum_{t=1}^T \ell_{it}^{\alpha_i\alpha_i}(\theta, \alpha_i)$, and

$$\widehat{V}^{\alpha_i}(\theta, \alpha_i) = \frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^{\alpha_i}(\theta, \alpha_i)(\ell_{t-j}^{\alpha_i}(\theta, \alpha_i))',$$

where $m \to \infty$ is a bandwidth parameter that satisfies $m/T^{1/2} \to 0$, and $K(\cdot)$ is a kernel function with $K(0) = 1$. Similarly, if interest centers on the posterior mean, we have that

$$\pi_{iT}^{E}(\theta, \alpha_i) = |-\widehat{H}_T^{\alpha_i\alpha_i}|^{1/2} \exp\{\frac{1}{2}\text{trace}[(\widehat{H}_T^{\alpha_i\alpha_i}(\theta, \alpha_i))^{-1}\widehat{V}^{\alpha_i}(\theta, \alpha_i)]\} \tag{4.12}$$

satisfies (4.9) and (4.10).[11]

## 5. Examples

5.1. **Normal Mean and Variance.** Let the data $y$ be a $T$-vector $y \sim \mathcal{N}(\alpha\, \iota_T, \sigma^2 I_T)$, where $\alpha$ is a scalar mean and $\iota_T$ is a $T$-vector of ones. The log likelihood takes the form

$$L = -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y-\alpha)'(y-\alpha)$$

We consider two estimators of the parameter $\theta = (\sigma, \alpha)'$. Under a flat prior, the joint posterior mode is the ML estimate, $\widehat{\alpha} = \bar{y}$ and

$$\widehat{\sigma}_M^2 \quad = \quad \frac{1}{T}(y-\bar{y})'(y-\bar{y}) \quad = \quad \sigma^2\frac{\chi_{T-1}^2}{T}.$$

The posterior means can be obtained by completing the square and rewriting the likelihood[12]

$$L = -\frac{1}{2}\log\left(\frac{\sigma^2}{T}\right) - \frac{T}{2\sigma^2}(\alpha-\bar{y})^2 - \frac{T-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y-\bar{y})'(y-\bar{y}) + \text{const.}$$

Recognizing the first and second pairs of right hand side terms as kernels of a normal and inverse gamma distribution, we have the usual normal-inverse gamma posteriors

$$p\left(\alpha|\sigma^2, y\right) \propto \left(\frac{\sigma^2}{T}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(\alpha-\bar{y})^2}{\sigma^2/T}\right) \qquad p\left(\sigma^2|y\right) \propto \left(\sigma^2\right)^{-\frac{(T-1)}{2}} \exp\left(-\frac{\frac{1}{2}(y-\bar{y})'(y-\bar{y})}{\sigma^2}\right)$$

The posterior mean for $\alpha$ is again the ML estimator, $\bar{y}$. Using the formula for the mean of the inverse gamma distribution, we have[13]

$$\widehat{\sigma}_E^2 \quad = \quad \frac{1}{T-5}(y-\bar{y})'(y-\bar{y}) \quad = \quad \sigma^2\frac{\chi_{T-1}^2}{T-5}.$$

[11] As in the general case, the priors defined in (4.11) and (4.12) would need to be modified as in Appendix A in cases where $-\widehat{H}_T^{\alpha_i\alpha_i}(\theta, \alpha_i)$ does not have minimum eigenvalue bounded away from zero over the parameter space.

[12] Here and below, const. denotes a constant which does not involve the parameter vector $\theta$.

[13] The inverse gamma density can be written $p(x|a,b) \propto x^{-(a+1)}e^{-\frac{b}{x}}$, and has mean $E(x) = \frac{b}{a-1}$. The posterior for $x = \sigma^2$ has this form with $b = \frac{1}{2}(y-\bar{y})'(y-\bar{y})$, and $a+1 = \frac{T-1}{2}$.

Both estimators of $\alpha$ are unbiased, while the estimates of $\sigma^2$ are biased in different directions: $E\left(\widehat{\sigma}_M^2\right) = \frac{T-1}{T}\sigma^2$, and $E\left(\widehat{\sigma}_E^2\right) = \frac{T-1}{T-5}\sigma^2$. We can construct bias reducing priors for the posterior mode and mean by solving the differential equations given in Theorem 2. Again using $\widehat{H} = T^{-1}\left[\partial^2_{\theta\theta} L\right]$ and $H = E[\widehat{H}]$ to denote the Hessian and its expectation, and defining $\widetilde{\psi} = T^{-\frac{1}{2}} H^{-1}\left[\partial_\theta L\right]$, we have

$$\begin{aligned}
\frac{\partial \log \pi_M}{\partial \left(\sigma^2\alpha\right)'} &= -E\left[\sqrt{T}\widehat{H}\widetilde{\psi}\right] - \frac{1}{2}\frac{\partial H}{\partial \sigma} E\left[\widetilde{\psi}\widetilde{\psi}_\sigma\right] - \frac{1}{2}\frac{\partial H}{\partial \alpha} E\left[\widetilde{\psi}\widetilde{\psi}_\alpha\right] \\
\frac{\partial \log \pi_E}{\partial \left(\sigma^2\alpha\right)'} &= \frac{\partial \log \pi_M}{\partial \left(\sigma^2\alpha\right)'} + \frac{1}{2}\frac{\partial \log|-H|}{\partial \theta},
\end{aligned}$$

where we have used $\widetilde{\psi}_\sigma$ to denote the element of $\widetilde{\psi}$ corresponding to the $\sigma$ (the first element, since $\sigma$ is the first element of the parameter vector). In this example, we can solve these differential equations in closed form. After some algebra, we find[14]

$$\begin{aligned}
\frac{\partial \log \pi_M}{\partial \left(\sigma^2, \alpha\right)'} = \begin{pmatrix} \frac{2}{\sigma^2} + \frac{1}{\sigma^2} \\ 0 \end{pmatrix} + \begin{pmatrix} -\frac{2}{\sigma^2} \\ 0 \end{pmatrix} + \begin{pmatrix} -\frac{1}{2\sigma^2} \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2\sigma^2} \\ 0 \end{pmatrix} \quad &\Rightarrow \quad \log \pi_M = \frac{1}{2}\log \sigma^2 \\
\frac{\partial \log \pi_E}{\partial \left(\sigma^2, \alpha\right)'} = \begin{pmatrix} \frac{1}{2\sigma^2} \\ 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} -\frac{4}{\sigma^2} - \frac{1}{\sigma^2} \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{2}{\sigma^2} \\ 0 \end{pmatrix} \quad &\Rightarrow \quad \log \pi_E = -2\log \sigma^2
\end{aligned}$$

By Theorem 2, $\pi_M$ removes the first order bias in the joint posterior mode while $\pi_E$ removes the first order bias in the marginal posterior means. Note that neither prior involves $\alpha$—both estimators of the location parameter are already unbiased and do not change under these priors.

Under the first prior $\pi_M$, the first term in the log posterior $L + \log \pi_M$ becomes $-\frac{(T-1)}{2}\log \sigma^2$. Differentiating the joint posterior to find the joint mode, the resulting estimate of $\sigma^2$ is identical to $\widehat{\sigma}_M^2$ but with $T-1$ instead of $T$ in the denominator, and is exactly unbiased. Using the second prior $\pi_E$, the first term in the marginal posterior $p(\sigma^2|y)$ becomes $(\sigma^2)^{-\frac{(T+3)}{2}}$. Again using the formula for the mean of an inverse gamma variate, the resulting marginal posterior mean for $\sigma^2$ is also unbiased.

5.1.1. *The Neyman-Scott problem.* We consider a classic generalization of the previous example where the data are normally distributed with group specific means. For simplicity

[14] As when we rewrote the likelihood, we may omit the constants of integration in the solutions for $\log \pi_M$ and $\log \pi_E$ because they do not involve the parameters.

suppose the data are $\{y_i\}_{i=1}^n$, where each $y_i$ is a $T$-vector and $y_i \sim \mathcal{N}(\alpha_i \iota_T, \sigma^2 I_T)$. The log likelihood can be written

$$L = -\frac{nT}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i (y_i - \alpha_i)'(y_i - \alpha_i),$$

or, completing the square and ignoring constants as before,

$$L = \sum_i \left[-\frac{1}{2}\log\left(\frac{\sigma^2}{T}\right) - \frac{T}{2\sigma^2}(\alpha_i - \bar{y}_i)^2\right] + \left[-\frac{n(T-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i (y_i - \bar{y}_i)'(y_i - \bar{y}_i)\right].$$

Again the posterior distribution for $\sigma^2$ is inverse gamma, while the posteriors for $\alpha_i|\sigma^2, y$ are independent normal with means $\bar{y}_i$ and variance $\frac{\sigma^2}{T}$. Under flat priors, the joint posterior mode and marginal posterior mean give the same estimates for the location parameters, $\widehat{\alpha}_i = \bar{y}_i$.

As before, the joint posterior mode and marginal posterior mean yield biased estimates of $\sigma^2$. The joint posterior mode is the maximum likelihood estimator,

$$\widehat{\sigma}^2_M \quad = \quad \frac{1}{nT}\sum_i (y_i - \bar{y}_i)'(y_i - \bar{y}_i) \quad = \quad \sigma^2 \frac{1}{n}\sum_i \frac{\chi^2_{T-1}}{T},$$

while the marginal posterior mean $\widehat{\sigma}^2_E$ is

$$\widehat{\sigma}^2_E \quad = \quad \frac{1}{n(T-1)-4}\sum_i (y_i - \bar{y}_i)'(y_i - \bar{y}_i) \quad = \quad \sigma^2 \frac{1}{n}\sum_i \frac{\chi^2_{T-1}}{(T-1)-\frac{4}{n}}.$$

Solving for the bias correcting priors is similar to the $n = 1$ case. Here the differential equations become

$$\frac{\partial \log \pi_M}{\partial(\sigma^2, \alpha)'} = \begin{pmatrix} \frac{2}{\sigma^2} + \frac{n}{\sigma^2} \\ 0_{n\times 1} \end{pmatrix} + \begin{pmatrix} -\frac{2}{\sigma^2} \\ 0_{n\times 1} \end{pmatrix} + \begin{pmatrix} -\frac{n}{2\sigma^2} \\ 0_{n\times 1} \end{pmatrix} = \begin{pmatrix} \frac{n}{2\sigma^2} \\ 0_{n\times 1} \end{pmatrix} \quad \Rightarrow \quad \log \pi_M = \frac{n}{2}\log\sigma^2$$

$$\frac{\partial \log \pi_E}{\partial(\sigma^2, \alpha)'} = \begin{pmatrix} \frac{n}{2\sigma^2} \\ 0_{n\times 1} \end{pmatrix} + \frac{1}{2}\begin{pmatrix} -\frac{4}{\sigma^2} - \frac{n}{\sigma^2} \\ 0_{n\times 1} \end{pmatrix} = \begin{pmatrix} -\frac{2}{\sigma^2} \\ 0_{n\times 1} \end{pmatrix} \quad \Rightarrow \quad \log \pi_E = -2\log\sigma^2$$

Using the same argument as in the $n = 1$ case, we can verify that the posterior mode under $\pi_M$ and the posterior mean under $\pi_E$ are exactly unbiased estimators for $\sigma^2$.

Neyman and Scott (1948) use this example as a cautionary tale. Taking expectations, we have $E(\widehat{\sigma}^2_M) = \frac{T-1}{T}\sigma^2$ and $E(\widehat{\sigma}^2_E) = \frac{T-1}{(T-1)-\frac{4}{n}}\sigma^2$. If we think of $T$ as fixed while the number of groups, $n$, increases, the maximum likelihood estimator of $\sigma^2$ is inconsistent. Interestingly,

the marginal posterior mean $\widehat{\sigma}_E^2$ is consistent for $\sigma^2$ with $T$ fixed.[15] We can also see this in the two bias correcting priors, as $\pi_M$ depends on $n$, while $\pi_E$ does not.

As we noted in Section 4.3, when working in panel settings with group specific parameters, one may wish to simplify the expressions for $\pi_M$ and $\pi_E$ by neglecting terms that vanish when $T$ is fixed and $n$ goes to infinity. In this example, the inverse information matrix has the following block structure:

$$-H^{-1} = \frac{1}{T}\begin{bmatrix} \frac{2}{n}\sigma^4 & 0_{1\times n} \\ 0_{n\times 1} & \sigma^2 I_n \end{bmatrix} \approx \frac{1}{T}\begin{bmatrix} 0 & 0_{1\times n} \\ 0_{n\times 1} & \sigma^2 I_n \end{bmatrix},$$

with the last equality approximate when $n$ is large relative to $T$. If we use this approximation in place of $H^{-1}$ in the derivation of $\frac{\partial \log \pi_M}{\partial \theta}$ and $\frac{\partial \log \pi_E}{\partial \theta}$ for this example, the algebra is simplified considerably, and the end result is that the terms which do not have $n$ in the numerator drop out. While this does not affect the solution for $\pi_M$, under this approximation the solution for $\pi_E$ is a constant (i.e., a flat prior). This may seem unfortunate, as using the full solution for $\pi_E$ yields an exactly unbiased estimate of $\sigma^2$. However, recall that $E\left(\frac{\widehat{\sigma}_E^2}{\sigma^2}\right) = \frac{T-1}{(T-1)-\frac{4}{n}}$, so when $n$ is large the bias in $\widehat{\sigma}_E^2$ is very small.

A similar approximation is very useful when working with the data dependent approximations to $\pi_M$ and $\pi_E$. In this example, this amounts to setting the first row and column of the sample information matrix to zero. The resulting data dependent approximation to $\pi_M$ is

$$-\widehat{H}^{-1} \approx \frac{1}{T}\begin{bmatrix} 0 & 0_{1\times n} \\ 0_{n\times 1} & \sigma^2 I_n \end{bmatrix} \quad \Rightarrow \quad \log \widetilde{\pi}_M \approx \frac{-1}{2T\sigma^2}\sum_i (y_i - \alpha_i)'(y_i - \alpha_i)$$

Using this as a pseudo-prior results in a log posterior

$$L + \log \tilde{\pi}_M = -\frac{nT}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\left(\frac{T+1}{T}\right)\sum_i (y_i - \alpha_i)'(y_i - \alpha_i)$$

The resulting posterior mode for $\sigma^2$ is then simply $\left(\frac{T+1}{T}\right)\widehat{\sigma}_M^2$, which has expectation

$$\left(\frac{T+1}{T}\right)\left(\frac{T-1}{T}\right)\sigma^2 = \left(1 - \frac{1}{T^2}\right)\sigma^2.$$

[15] This is a fortunate accident resulting from the structure of the log likelihood in the linear model. In nonlinear panel models, simply replacing the joint mode with the marginal posterior mean does not result in a 'fixed-$T$ consistent' estimate of the common parameters.

While the estimator is still not consistent with $T$ fixed, the order of bias has been reduced from $T^{-1}$ to $T^{-2}$, which can be a substantial improvement when $T$ is small.

## 6. Monte Carlo Study

In this section, we explore the finite-sample frequentist properties of the posterior modes and means that result from the use of the bias-reducing priors through a brief simulation study. We consider the case of estimating a simple dynamic model defined by

$$y_{it} = \alpha_i + \rho y_{it-1} + u_{it} \tag{6.1}$$

for $i = 1, ..., N$ and $t = 1, ..., T$ where the $u_{it}$ are iid normal innovations with mean 0 and variance $\sigma^2$.

In the $N = 1$ case, the model is the conventional univariate AR(1) model which has been much studied and is commonly used in applications. With $N > 1$, the model is the standard dynamic panel model without covariates. In this case, it is well known that the usual maximum likelihood estimates of $\rho$ are inconsistent with $T$ fixed and $N \to \infty$ and may be severely biased for short $T$.[16] While our results are derived under asymptotics where $T \to \infty$ with $N$ fixed, the simulation results show that our approach may dramatically improve estimation even when $T$ is small. Our approach is also quite similar to the approach considered in Hahn and Kuersteiner (2002), Hahn and Newey (2004), and Hahn and Kuersteiner (2004) and related penalty function methods[17] which produce estimators of common parameters in panel data models with incidental parameters that have asymptotically normal sampling distributions that are centered on the true parameter value in asymptotics where $N \to \infty$, $T \to \infty$, and $N/T = O(1)$. In the simple dynamic panel context, $\sqrt{N}$ consistent GMM estimators of $\rho$ are available; see, for example, Arellano (2003). However, GMM approaches rely on differencing methods and are not readily generalized to nonlinear models. Lancaster (2002) considers a Bayesian approach based upon using a transformation of the fixed effects which makes them information orthogonal to the remaining parameters and then integrating them out of the model and produces a $\sqrt{N}$ consistent estimator of $\rho$ and $\sigma^2$ in the dynamic panel case. Our "closed form" prior for the posterior mean is equivalent to the Jacobian

[16] See for example Nickell (1981).

[17] See, for example, Arellano and Hahn (2005) and Bester and Hansen (2005).

of the transformation implied by Lancaster's (2002) approach in the dynamic panel model excepting a term which is O(1/NT) and so will also produce $\sqrt{N}$ consistent estimates of the common parameters in this context.[18] However, our approach requires neither orthogonalization or integration and so may be simpler to implement in practice in more complicated settings.

In our simulations, we consider cases with $N = 1$ and $T \in \{50, 200\}$ and $N = 100$ and $T \in \{4, 12\}$. For all sample sizes we consider two values of $\rho$, .6 and .9, and set $\sigma^2 = 1$. When $N = 1$, we use $\alpha = 0$ and draw $y_0$ from a $N(0, \frac{1}{1-\rho^2})$ at each iteration. Otherwise, the $\alpha_i$ are drawn as iid $N(0, 1)$ random variables at each iteration, and the $y_{i0}$ are then drawn as $N(\frac{\alpha_i}{1-\rho}, \frac{1}{1-\rho^2})$ variables. In all cases, we estimate $N + 2$ parameters, $\rho$, the $\alpha_i$, and $\sigma^2$, at each iteration and use 1000 Monte Carlo iterations. We estimate the model parameters using both joint posterior modes and posterior means. When considering the joint posterior mode, we obtain estimates under a flat prior ($\pi = 1$) which correspond to the MLE, under the data-dependent prior designed to remove bias from the posterior mode defined in equation (4.3), and under the closed form version of the prior satisfying the differential equations given in (4.1).[19] When considering the posterior mean, we obtain estimates under a flat prior ($\pi = 1$), under the data-dependent prior designed to remove bias from the posterior mean defined in equation (4.5), and under the closed form version of the prior satisfying the differential equations given in (4.2).[20] We also consider two estimators that remove bias by directly correcting the MLE by subtracting an estimate of the higher order bias obtained in (3.9), that is estimators of the form

$$\widetilde{\theta} = \widehat{\theta}^M - \widehat{B}(\widehat{\theta}^M)/T$$

[18]It can also be shown that the "closed form" prior for the posterior mode will produce a $\sqrt{N}$ consistent estimator of $\rho$ and $\sigma^2$. See Bester and Hansen (2005).

[19]The form of the closed form prior is given in equation (D.15) in Appendix D. In the panel case, we make use of the panel simplification (4.11) when using the data dependent prior as this substantially eases computation.

[20]The form of the closed form prior is given in equation (D.16) in Appendix D. In the panel case, we make use of the panel simplification (4.12) when using the data dependent prior as this substantially eases computation. It is also interesting to note that the prior defined in (D.16) corresponds to the "matching prior" in the sense that as $T \to \infty$ the difference between the two priors goes to 0. For early work on matching priors, see Welch and Peers (1963).

where $\widehat{\theta}^M$ is the maximum likelihood estimate of $\theta$ and $\widehat{B}(\theta)$ is an estimate of (3.9). As with the Bayesian estimators, we consider both a data-dependent approximation to the bias given in (3.9) and a closed form version that explicitly computes the expectations.[21] Posterior modes are obtained by numerically maximizing the corresponding likelihood and posterior means are obtained via Markov Chain Monte Carlo.[22] Finally, all of the data dependent approaches rely on the choice of a truncation parameter, $M$. In the panel cases, $M = 1$ was a natural choice due to the short $T$. In the time series cases, we chose $M = 2$ with $T = 50$ and $M = 3$ with $T = 200$.

Simulation results are summarized in Tables 1 and 2. Table 1 contains results for the time series ($N = 1$) case, and Table 2 contains results for the panel ($N = 100$) case. In both tables, rows labeled $\widehat{\rho}^M$ give results for the MLE, and rows labeled $\widehat{\rho}^M_{DD}$ and $\widehat{\rho}^M_{CF}$ give results for the joint posterior mode based on the data-dependent and closed form priors, respectively. Similarly, rows labeled $\widehat{\rho}^E$ give results for the posterior mean under a flat prior, and $\widehat{\rho}^E_{DD}$ and $\widehat{\rho}^E_{CF}$ give results for the posterior means based on the data-dependent and closed form priors, respectively. Finally, the rows labeled $\widehat{\rho}^M_{BCDD}$ and $\widehat{\rho}^M_{BCCF}$ correspond to the bias-corrected maximum likelihood estimators defined above where $\widehat{\rho}^M_{BCDD}$ approximates the bias using sample averages and $\widehat{\rho}^M_{BCCF}$ explicitly computes the bias in terms of the model parameters and then plugs in the ML estimates of these parameters to estimate the bias. We report only results for $\rho$ which is typically the parameter of interest, though results for other parameters are available from the authors.

For each estimator and data configuration, we report the bias of the estimator along with the root mean squared error (RMSE) and the mean absolute deviation (MAD). We also report frequentist coverage probabilities of 95% level interval estimates. For $\widehat{\rho}^M$, $\widehat{\rho}^M_{DD}$, $\widehat{\rho}^M_{CF}$, $\widehat{\rho}^M_{BCDD}$, and $\widehat{\rho}^M_{BCCF}$, these intervals are calculated from the asymptotic distribution where the asymptotic variance is estimated using the usual formula and the estimate of $\sigma^2$ corresponding to the appropriate estimate of $\rho$. For $\widehat{\rho}^E$, $\widehat{\rho}^E_{DD}$, and $\widehat{\rho}^{CF}$, the interval estimate corresponds to a 95% level Bayesian credibility interval obtained from the MCMC draws.

[21] In the time series case, the closed form expression is given in Shaman and Stine (1988). For the panel context, we make use of the panel simplification suggested in Section 4.3. In this case the data dependent version corresponds to the correction suggested in Hahn and Kuersteiner (2004) and the closed form version to the correction in Hahn and Kuersteiner (2002).

[22] Details about the MCMC sampler are available from the authors upon request.

Looking first at Table 1, we see that the bias-reducing priors are quite effective at removing bias in all cases with the closed form version of the priors being the most effective. Indeed using the closed form version of the priors results in an approximately unbiased estimator in all four cases. Results with the data-dependent prior are more varied, ranging between a 20% reduction in the bias when $T = 50$ and $\rho = .9$ to almost completely removing the bias when $T = 200$ and $\rho = .6$. The closed form priors also do better in terms of bias than do either of the bias-corrected ML estimators, and the biases resulting when the data-dependent priors are used are similar to those obtained from the bias-corrected ML procedures. As would be expected, these reductions in bias do come at the cost of increased variance which can be seen in the RMSE and MAD of the estimators. Indeed, with $T = 50$ and $\rho = .6$, the estimators resulting from the use of the closed form priors perform worse in terms of RMSE or MAD than do any of the other bias-reduced estimators or even the uncorrected MLE. In all cases, the estimators resulting from the use of the data-dependent priors perform quite favorably in terms of RMSE and MAD and perform similarly to the bias-reduced ML based upon a data-dependent estimate of the bias. In terms of coverage probabilities, all estimators perform reasonably well when $\rho = .6$ with both $T = 50$ and $T = 200$. When $\rho = .9$ and $T = 50$, there are substantial distortions in the coverage probabilities of both the MLE and the posterior mean based on a flat prior. These distortions are essentially removed when one considers interval estimates obtained using the closed form bias-reducing priors though the data-dependent versions do not seem to improve the coverage probabilities substantially. Overall, the evidence from this case suggests that the priors are effective at doing what they were designed for, removing bias. However, this bias-reduction does come at the cost of increased variance. In most cases, the reduction in bias appears to dominate the increase in variance in that RMSE and MAD tend to also decrease relative to uncorrected estimates. The Bayesian estimators also compare favorably to more standard bias-corrected frequentist maximum likelihood estimates.

The results from Table 2 are much less ambiguous. All of the bias reductions substantially reduce the bias relative to the uncorrected maximum likelihood estimator, and since bias is the dominant component of both the RMSE and MAD, they also substantially reduce the RMSE and MAD. Thus, the bias-reduced estimators clearly dominate the uncorrected ML or the posterior mean based on a flat prior. The performance of the Bayesian estimators when

the data-dependent priors are used is similar to the performance of the bias-reduced maximum likelihood when the bias is estimated using sample averages. However, the Bayesian estimators resulting from the use of the closed-form priors clearly dominate the bias-reduced maximum likelihood estimator which explicitly calculates the higher-order bias of the estimator. The dominance is not surprising in this case, since as discussed above, the Bayesian estimators based on the closed-form priors are $\sqrt{N}-$ consistent in this case while the bias-reduced ML is not. However, we do not believe that this dominance would hold in other models; and in general, one would expect their performance to be similar.

The simulation results confirm that the bias-reducing priors are indeed effective in reducing the bias of the corresponding posterior mode or mean. In the majority of cases considered, this reduction in bias also translates into a reduction in RMSE or MAD. This improvement is especially evident in the panel case where bias is the dominant feature of the sampling distribution of the maximum likelihood estimator. The use of these or other objective Bayes priors also seems desirable in panel situations as elicitation of a prior over a high-dimensional parameter space may be quite difficult. Overall, we feel the simulation results suggest that the bias-reducing priors may be quite useful in some settings, especially in settings where bias is likely to play a large role and elicitation of subjective priors is difficult.

## 7. Conclusion

We propose a set of objective Bayesian priors that remove the first order frequentist bias in the posterior mode and posterior mean. These priors are based on asymptotic expansions for the two estimators, which we develop rigorously under mild regularity conditions. Our approach therefore applies to a very general class of likelihood models.

Similar to the matching priors of Welch and Peers (1963), Ghosh and Mukerjee (1992), and Mukerjee and Dey (1993), our priors are defined implicitly as the solution to a set of differential equations. Although these differential equations involve only the scores and their derivatives, solutions will not always be available. We therefore present a simple data-dependent approximation to the prior that also removes first order bias asymptotically. The data dependent approximation involves only the sample information matrix and outer product of scores, both of which are widely used by practioners for frequentist inference.

We illustrate our approach in two example models, including the classical problem of Neyman and Scott (1948), where both the prior and data-dependent approximation may be derived explicitly and interpreted. We also present a brief Monte Carlo study using a first order autoregressive model in scalar time series and panel settings. The results suggest that our bias correction may be quite useful in time series and panel data applications.

## Appendix A. Data-Dependent Bias Reducing Priors: The General Case

In this appendix, we present generalizations of the data-dependent priors introduced in Section 4.2 that accommodate models in which the log-likelihood is not strictly concave. The difficulty that arises in these settings is that the Hessian term that appears in the expressions for the data-dependent bias-reducing priors for the posterior mode and mean, (4.3) and (4.5), will not be negative definite across the parameter space but will be negative definite near local maxima of the log-likelihood. This property of the likelihood will result in singularities in the $\widehat{H}_T(\theta)$ for some values of $\theta$ in the parameter space unless the parameter space is restricted a priori to rule such values of $\theta$ out. As an alternative to restricting the parameter space, we consider replacing the Hessian term in the data-dependent priors with a smoothed approximation that is by construction negative definite where we allow the degree of smoothing to depend on the sample size. The replacement makes the priors more well-behaved by insuring that the priors do not equal 0 or infinity within the parameter space but does not affect the asymptotic properties of the estimator near $\theta_0$ under Assumption 1 so retains the bias-reducing properties of priors (4.3) and (4.5).

Specifically, we consider replacing $-\widehat{H}_T(\theta)$ in equations (4.3) and (4.5) with

$$\widehat{\Gamma}_T(\theta) = -\widehat{H}_T(\theta)k_T + A_T(\theta)(1 - k_T) \tag{A.1}$$

to obtain

$$\widetilde{\pi}_T^M(\theta) = \exp\{\frac{1}{2}\text{tr}\left[-\widehat{\Gamma}_T(\theta)^{-1}\widehat{V}(\theta)\right]\} \tag{A.2}$$

and

$$\widetilde{\pi}_T^E(\theta) = |\widehat{\Gamma}_T(\theta)|^{1/2}\exp\{\frac{1}{2}\text{tr}\left[-\widehat{\Gamma}_T(\theta)^{-1}\widehat{V}(\theta)\right]\} \tag{A.3}$$

where

$$k_T = K(-\widehat{\lambda}_{\max}, \delta_T, h_{\delta_T})(1 - K(-\widehat{\lambda}_{\min}, \Delta_T, h_{\Delta_T})) \tag{A.4}$$

for $\widehat{\lambda}_{\min}$ and $\widehat{\lambda}_{\max}$ the minimum and maximum eigenvalues of $\widehat{H}_T(\theta)$ respectively, $\delta_T > 0$, $h_{\delta_T} > 0$, $\Delta_T > 0$, $h_{\Delta_T} > 0$, $\delta_T \to 0$, $h_{\delta_T} \to 0$, $\Delta_T \to \infty$, and

$$K(\lambda, \delta, h) = \left( -\frac{1}{24} + \frac{13}{12}\frac{\lambda - \delta}{h} - \frac{1}{24}\left(\frac{2(\lambda - \delta)}{h} - 1\right)^{13} \right) 1(\delta < \lambda < \delta + h) + 1(\lambda \geq \delta + h)$$

and $A_T(\theta)$ is a uniformly positive definite matrix in $T$ and $\theta$ with minimum eigenvalue uniformly greater than or equal to $\delta_T$ and maximum eigenvalue uniformly less than or equal to $\Delta_T + h_{\Delta_T}$.[23] It follows by construction that $\widehat{\Gamma}_T(\theta)$ is uniformly positive definite. Also, since $H_T(\theta)$ is negative definite with bounded eigenvalues when evaluated at $\theta_0$, it follows that if $\delta_T$ and $h_T$ approach 0 and $\Delta_T$ approaches infinity slowly enough as the sample size increases, the truncation will not affect the asymptotic properties of the estimators constructed using the truncated data-dependent priors.

In order to verify that the data-dependent prior which replaces $-\widehat{H}_T(\theta)$ with $\widehat{\Gamma}_T(\theta)$ removes the higher-order bias from the resulting estimators, we further strengthen the regularity conditions summarized in Assumptions 1 and 2.

**Assumption 3.** *For the model defined in Section 2 and for $\widehat{\Gamma}_T(\theta)$ given in (A.1), suppose that (i) $A_T(\theta)$ is a uniformly positive definite matrix in $T$ and $\theta$ with minimum eigenvalue uniformly greater than or equal to $\delta_T$ and maximum eigenvalue uniformly less than or equal to $\Delta_T + h_{\Delta_T}$ and $\Delta^j A_T(\theta)$ exists and has bounded eigenvalues for all $\theta \in \Theta$ and $0 \leq j \leq 6$; (ii) $\delta_T > 0$, $h_{\delta_T} > 0$, $\Delta_T > 0$, $h_{\Delta_T} > 0$; (iii) for $0 < \epsilon < 1/2$, $m = o(T^{1/2-\epsilon})$ for $m$ the bandwidth used in defining $\widehat{V}(\theta)$ and $\delta_T = O(T^{-\epsilon/7})$; also, $h_{\delta_T} = O(T^{-\alpha})$ for some $\alpha > 0$, $\Delta_T = O(T^{\alpha})$ for some $\alpha > 0$, and $h_{\Delta_T} = O(T^{\alpha})$ for some $\alpha \geq 0$; (iv) the maximum and minimum eigenvalue of $\widehat{H}_T(\theta)$ are simple; (v) $\widehat{V}_T(\theta)$ is positive semi-definite for all $\theta \in \Theta$; (vi) $\{w_t,\ t = 1, 2, ...\}$ is a mixing sequence that satisfies $\alpha(m) = O(m^{\frac{-3r}{r-2}-\epsilon})$ for some $\epsilon > 0$ and $r > 2$; (vii) there exists a function $M_t(w_t)$ such that for $0 \leq j \leq 8$, all $w_t \in \mathcal{W}_T = \mathcal{Y}_t \times \mathcal{X}_t$, and all $\theta \in \mathcal{G}$ where $\mathcal{G}$ is an open, convex set containing $\Theta$, $\Delta^j \ell_t(\theta)$ exists, $\sup_{\theta \in \mathcal{G}} \|\Delta^j \ell_t(\theta)\| \leq M_t(w_t)$, and $\sup_t \mathrm{E}\|M_t(w_t)\|^{7r+\delta} \leq M < \infty$ for some $\delta > 0$.*

Assumption 3 is similar to Assumption 2 which is sufficient in the case where the likelihood is strictly concave. The chief differences arise in the rate conditions on the truncation parameters imposed in Conditions (i)-(iii) of Assumption 3. These rate conditions are sufficient to insure that the degree of truncation of the Hessian diminishes slowly enough with the sample size to guarantee that the posterior using the data-dependent prior is uniformly well-behaved without affecting the

[23] Note that any six-times continuously differentiable weighting function $k_T$ that goes to one slowly enough as the sample size increases in such a way as to bound the minimum and maximum eigenvalues of $\widehat{\Gamma}_T(\theta)$ away from 0 and infinity could also be used. We could also allow for $A_T(\theta)$ to be estimated or otherwise depend on the data as long as the resulting $A_T(\theta)$ is six-times continuously differentiable and $\mathrm{E}\|\Delta^j A_T(\theta)\|^{7r+\delta} < \infty$ for $0 \leq j \leq 6$.

bias-reducing properties of the estimators. Condition (iv) imposes that the maximum and minimum eigenvalues of $\widehat{H}_T(\theta)$ are unique which guarantees that they are smooth functions of the elements of $\widehat{H}_T(\theta)$ and does not seem restrictive for most applications. The other difference arises in that Assumption 3 imposes stronger moment conditions than Assumption 2 due to the fact that the truncation depends on the data through $\lambda_{\min}$ and $\lambda_{\max}$.

As above, we may now state a general version of Theorem 3.

**Theorem 4.** *If $T \to \infty$ and $m \to \infty$ such that Assumption 3 holds, Conditions (vii)-(ix) of Assumption 1 are satisfied by $\log[\widetilde{\pi}_T^M(\theta)]$ and $\log[\widetilde{\pi}_T^E(\theta)]$ and*

$$\frac{\partial}{\partial\theta}\log[\widetilde{\pi}_T^M(\theta)] - \left(-\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \frac{1}{2}\sum_{j=1}^{k}\mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j\right) = o_p(1)$$

*and*

$$\frac{\partial}{\partial\theta}\log[\widetilde{\pi}_T^E(\theta)] - \left(-\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \sum_{j=1}^{k}\mathrm{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j\right) = o_p(1).$$

*It follows that $\widetilde{\pi}_T^M(\theta)$ removes the higher-order bias from the posterior mode and that $\widetilde{\pi}_T^E(\theta)$ removes the higher-order bias from the posterior mean.*

## Appendix B. Preliminary Lemmas

Consistency and asymptotic normality of the posterior mode are important prerequisites for the higher order expansions provided in Theorem 1. We briefly verify consistency and asymptotic normality of the posterior mode in Lemmas B.1 and B.2 below.

**Lemma B.1.** *For $\widehat{\theta}^M = \arg\max_\theta T^{-1}(\ell_T(\theta) + \gamma_T(\theta))$, $\widehat{\theta}^M \xrightarrow{p} \theta_0$ if Assumption 1 is satisfied.*

**Proof.** The proof follows by a simple modification of a conventional consistency proof, e.g. the proof of Theorem 4.2 in Wooldridge (1994). Let $Q_T(\theta) = T^{-1}(\ell_T(\theta) + \gamma_T(\theta))$ and $\bar{Q}_T(\theta) = T^{-1}\sum_{t=1}^{T}\mathrm{E}[\ell_t(\theta)]$. Then

$$\begin{aligned}
\mathrm{P}\left(\sup_{\theta\in\Theta}|Q_T(\theta) - \bar{Q}_T(\theta)| > \epsilon\right) &\leq \mathrm{P}\left(\max_{1\leq j\leq K}\sup_{\theta\in\zeta_j}|Q_T(\theta) - \bar{Q}_T(\theta)| > \epsilon\right) \\
&\leq \sum_j \mathrm{P}\left(\sup_{\theta\in\zeta_j}|Q_T(\theta) - \bar{Q}_T(\theta)| > \epsilon\right) \\
&\leq \sum_j \mathrm{P}\left(\sup_{\theta\in\zeta_j}\left(|T^{-1}\sum_{t=1}^{T}\ell_t(\theta) - \bar{Q}_T(\theta)| + |T^{-1}\gamma_T(\theta)|\right) > \epsilon\right)
\end{aligned}$$

where $\zeta_j = \zeta_\delta(\theta_j)$ is a ball of radius $\delta$ about $\theta_j$ and $\zeta_\delta(\theta_j)$ for $j = 1, ..., K(\delta)$ is a finite covering of $\Theta$ which exists since $\Theta$ is compact by assumption. Then, as in Wooldridge (1994), choose $\delta < 1$ such that

$$\begin{aligned} &\mathrm{P}\left(\sup_{\theta\in\zeta_j}\ (|T^{-1}\sum_{t=1}^{T}\ell_t(\theta) - \bar{Q}_T(\theta)| + |T^{-1}\gamma_T(\theta)|) > \epsilon\right) \qquad \text{(B.1)} \\ &\qquad \leq \mathrm{P}\left(|T^{-1}\sum_{t=1}^{T}(M_t - \mathrm{E}[M_t])| + |T^{-1}\sum_{t=1}^{T}(\ell_t(\theta_j) - \mathrm{E}[\ell_t(\theta_j)])| + \sup_{\theta\in\zeta_j}|T^{-1}\gamma_T(\theta)| > \epsilon/2\right) \end{aligned}$$

Then since $T^{-1}\sum_{t=1}^{T}(M_t - \mathrm{E}[M_t]) \xrightarrow{p} 0$, $T^{-1}\sum_{t=1}^{T}(\ell_t(\theta_j) - \mathrm{E}[\ell_t(\theta_j)]) \xrightarrow{p} 0$, and $\sup_{\theta\in\zeta_j} T^{-1}\gamma_T(\theta) \xrightarrow{p} 0$, there is a $T_0$ such that for all $T > T_0$ and $j$, the right-hand side of equation (B.1) is less than or equal to $\epsilon/K$. It then follows that $Q_T(\theta)$ satisfies a uniform weak law of large numbers and the consistency of $\widehat{\theta}^M$ for $\theta_0$ follows by the usual argument. ■

**Lemma B.2.** *Under Assumption 1,* $\sqrt{T}(\widehat{\theta}^M - \theta_0 \xrightarrow{d} N(0, -H_0^{-1})$.

**Proof.** The result follows from the usual argument. We have that $\widehat{\theta}^M$ satisfies

$$\begin{aligned} 0 &= T^{-1}\sum_{t=1}^{T}\ell_t^{\theta)(\widehat{\theta}^M} + T^{-1}\gamma_T^{\theta}(\widehat{\theta}^M) \\ &= T^{-1}\sum_{t=1}^{T}\ell_t^{\theta}(\theta_0) + T^{-1}\sum_{t=1}^{T}\ell_t^{\theta\theta}(\bar{\theta})(\widehat{\theta}^M - \theta_0) + T^{-1}\gamma_T^{\theta}(\widehat{\theta}^M) \end{aligned}$$

where $\bar{\theta}$ is an intermediate value between $\widehat{\theta}^M$ and $\theta_0$. It follows that

$$\sqrt{T}(\widehat{\theta}^M - \theta_0) = (-T^{-1}\sum_{t=1}^{T}\ell_t^{\theta\theta}(\bar{\theta}))^{-1}(T^{-1/2}\sum_{t=1}^{T}\ell_t^{\theta}(\theta_0) + T^{-1/2}\gamma_T^{\theta}(\widehat{\theta}^M)).$$

Then, under the conditions of Assumption 1, $T^{-1/2}\gamma_T^{\theta}(\widehat{\theta}^M) = o_p(1)$, $-T^{-1}\sum_{t=1}^{T}\ell_t^{\theta\theta}(\bar{\theta}) \xrightarrow{p} -H_0$, and $T^{-1/2}\sum_{t=1}^{T}\ell_t^{\theta} \xrightarrow{d} N(0, -H_0)$ by an appropriate CLT (e.g. White (2001) Theorem 5.20) and the Cramer-Wold device and noting that under the assumed dynamic completeness $\mathrm{Var}(T^{-1/2}\sum_{t=1}^{T}\ell_t^{\theta}) = -H_T$. The conclusion then follows immediately. ■

## Appendix C. Proofs

In this section, we prove Theorems 1 and 4. Theorem 2 is immediate from Theorem 1 under the hypotheses of the theorem, and the proof of Theorem 3 is quite similar to, but simpler than, the proof of Theorem 4. As such, proofs of Theorems 2 and 3 are omitted for brevity.

C.1. **Proof of Theorem 1.** We begin by considering the expansion of the posterior mean $\widehat{\theta}^M$. Using the same arguments to derive (3.5), we have

$$\widehat{\theta}^M - \theta_0 = T^{-1/2}\psi_1 - T^{-1}(H_T)^{-1}\bar{\gamma}_T^{\theta} - T^{-1}(H_T)^{-1}\left[\sqrt{T}\left(\widehat{H}_T - H_T\right)\right]\psi_1 - \frac{1}{2}T^{-1}\sum_{j=1}^{k}(H_T)^{-1}H_T^{\theta_j}\psi_1\psi_{j1} + R_T$$

where

$$\begin{aligned} R_T = -H_T^{-1}\Bigg(&\frac{1}{6}\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{H}_T^{\theta_j\theta_k}(\bar{\theta})(\widehat{\theta}^M - \theta_0)(\widehat{\theta}_j^M - \theta_{0j})(\widehat{\theta}_k^M - \theta_{0k}) + T^{-1}\gamma_T^{\theta\theta}(\bar{\theta})(\widehat{\theta}^M - \theta_0) \\ &+ (\widehat{H}_T - H_T)(\widehat{\theta}^M - \theta_0 - T^{-1/2}\psi_1) + T^{-1}(\gamma_T^{\theta} - \bar{\gamma}_T^{\theta}) \\ &+ \frac{1}{2}\sum_{j=1}^{p}\widehat{H}_T^{\theta_j}[(\widehat{\theta}^M - \theta_0)(\widehat{\theta}_j^M - \theta_{0j}) - T^{-1/2}\psi_1 T^{-1/2}\psi_{1j}]\Bigg) \end{aligned} \tag{C.1}$$

with $\dim(\theta) = p$, $\bar{\theta}$ an intermediate value between $\widehat{\theta}^M$ and $\theta_0$, and $\widehat{H}_T^{\theta_j\theta_k}(\theta) = T^{-1}\sum_{t=1}^{T}\ell_t^{\theta\theta\theta_j\theta_k}(\theta)$. The result will then follow if $R_T = o_p(T^{-1})$.

Under conditions (i)-(iii) of Assumption 1, $\widehat{H}_T^{\theta_j\theta_k}(\theta)$ satisfies a uniform weak law of large numbers, so $\widehat{H}_T^{\theta_j\theta_k}(\theta) = O_p(1)$. Also, from Lemma B.2, $\widehat{\theta}^M - \theta_0 = O_p(T^{-1/2})$. It then follows that

$$\widehat{H}_T^{\theta_j\theta_k}(\bar{\theta})(\widehat{\theta}^M - \theta_0)(\widehat{\theta}_j^M - \theta_{0j})(\widehat{\theta}_k^M - \theta_{0k}) = O_p(1)\left[O_p(T^{-1/2})\right]^3 = o_p(T^{-1}). \tag{C.2}$$

Following the derivation used in the proof of Lemma B.2, we also have $\widehat{\theta} - \theta_0 = T^{-1/2}\psi_1 + R_1$ where $R_1 = o_p(T^{-1/2})$. This equality then implies that

$$\widehat{\theta}^M - \theta_0 - T^{-1/2}\psi_1 = R_1 = o_p(T^{-1/2}) \tag{C.3}$$

and

$$\begin{aligned} (\widehat{\theta}^M - \theta_0)(\widehat{\theta}_j^M - \theta_{0j}) - T^{-1/2}\psi_1 T^{-1/2}\psi_{1j} &= T^{-1/2}\psi_1 R_{1j} + R_1 T^{-1/2}\psi_{1j} + R_1 R_{1j} \\ &= o_p(T^{-1}). \end{aligned} \tag{C.4}$$

Then, since conditions (i)-(iii) of Assumption 1 are sufficient for the elements of $\sqrt{T}(\widehat{H}_T - H_T)$ to follow a central limit theorem, e.g. White (2001) Theorem 5.20, it follows that

$$(\widehat{H}_T - H_T)(\widehat{\theta}^M - \theta_0 - T^{-1/2}\psi_1) = o_p(T^{-1}) \tag{C.5}$$

using (C.3). Also, since conditions (i)-(iii) of Assumption 1 are sufficient for $\widehat{H}_T^{\theta_j}$ to be bounded in probability, (C.4) implies that

$$\widehat{H}_T^{\theta_j}[(\widehat{\theta}^M - \theta_0)(\widehat{\theta}_j^M - \theta_{0j}) - T^{-1/2}\psi_1 T^{-1/2}\psi_{1j}] = o_p(T^{-1}). \tag{C.6}$$

Condition (ix) of Assumption 1 then implies that

$$T^{-1}(\gamma_T^{\theta} - \bar{\gamma}_T^{\theta}) = o_p(T^{-1}), \tag{C.7}$$

and condition (viii) of Assumption 1 gives that $T^{-1}\gamma_T^{\theta\theta}(\bar{\theta}) = o_p(T^{-1/2})$ from which

$$T^{-1}\gamma_T^{\theta\theta}(\bar{\theta})(\widehat{\theta}^M - \theta_0) = o_p(T^{-1}) \tag{C.8}$$

follows.

Plugging (C.2), (C.5), (C.6), (C.7), and (C.8) into the expression for the remainder given in (C.1) then yields the result for the posterior mode.

The argument for the posterior mean is similar. In particular, the conditions of Assumption 1 are sufficient for the conditions of Kass, Tierney, and Kadane (1990) Theorems 4 and 7 replacing $h_n(\theta)$ (in the notation of Kass, Tierney, and Kadane (1990)) with $-T^{-1}(\ell_T(\theta) + \gamma_T(\theta))$ (using our notation) from which the expansion

$$\widehat{\theta}^E = \widehat{\theta}^M + \frac{1}{2T}\left(Q_T^{\theta\theta}(\widehat{\theta}^M)\right)^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[\left(Q_T^{\theta\theta\theta_j}(\widehat{\theta}^M)\right)\left(Q_T^{\theta\theta}(\widehat{\theta}^M)\right)^{-1}\right] e_j + o_p(T^{-1}) \tag{C.9}$$

follows where $Q_T^{\theta\theta}(\theta) = \widehat{H}_T(\theta) + T^{-1}\gamma_T^{\theta\theta}(\theta)$ and $Q_T^{\theta\theta}(\theta) = \widehat{H}_T^{\theta_j}(\theta) + T^{-1}\gamma_T^{\theta\theta\theta_j}(\theta)$.

Now note that $T^{-1}\gamma_T^{\theta\theta}(\widehat{\theta}^M)$ and $T^{-1}\gamma_T(\theta\theta\theta_j)(\widehat{\theta}^M)$ converge in probability to 0 under condition (viii) of Assumption 1. Also, the conditions of Assumption 1 are sufficient for $\widehat{H}_T(\widehat{\theta}^M) \xrightarrow{p} H_0$ and $\widehat{H}_T^{\theta_j}(\widehat{\theta}^M) \xrightarrow{p} H_0^{\theta_j}$. It follows that

$$\begin{aligned}
&\left(Q_T^{\theta\theta}(\widehat{\theta}^M)^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[\left(Q_T^{\theta\theta\theta_j}(\widehat{\theta}^M)\right)\left(Q_T^{\theta\theta}(\widehat{\theta}^M)\right)^{-1}\right] e_j - H_T^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[H_T^{\theta_j} H_T^{-1}\right] e_j \right. \\
&\quad = \left(Q_T^{\theta\theta}(\widehat{\theta}^M)\right)^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[\left(Q_T^{\theta\theta\theta_j}(\widehat{\theta}^M)\right)\left(Q_T^{\theta\theta}(\widehat{\theta}^M)\right)^{-1}\right] e_j - H_0^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[H_0^{\theta_j} H_0^{-1}\right] e_j \\
&\qquad + H_0^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[H_0^{\theta_j} H_0^{-1}\right] e_j - H_T^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[H_T^{\theta_j} H_T^{-1}\right] e_j \\
&\quad = o_p(1).
\end{aligned}$$

Then adding and subtracting $H_T^{-1} \sum_{j=1}^{p} \operatorname{tr}\left[H_T^{\theta_j} H_T^{-1}\right] e_j$ from equation (C.9) and plugging in the expansion for $\widehat{\theta}^M$ from (3.5) derived above yields the result for the posterior mean. ■

C.2. **Proof of Theorem 4.** The theorem is proven by showing that the conditions of Assumption 3 imply conditions (vii)-(ix) of Assumption 1 and that the derivatives of the log-priors corresponds to the negative of the asymptotic bias. For brevity, we consider only $\widetilde{\pi}_T^M(\theta)$. The arguments for

$\widetilde{\pi}_T^E(\theta)$ are similar but somewhat more cumbersome due to the presence of the additional term $|\widehat{\Gamma}_T(\theta)|$.

For condition (vii) of Assumption 1, we note that $\widehat{V}_T(\theta)$ is positive semi-definite by assumption and by construction $\widehat{\Gamma}_T(\theta)$ is positive definite with bounded eigenvalues. It then follows that under the differentiability conditions in condition (vii) of Assumption 3, $\exp\{\frac{1}{2}\text{tr}\left[-\widehat{\Gamma}_T(\theta)^{-1}\widehat{V}_T(\theta)\right]\} > 0$ for all $T$ and $\theta$. It then follows immediately that $\widetilde{\pi}_T^M(\theta) > 0$.

For condition (ix) of Assumption 1, we differentiate $\log\left[\widetilde{\pi}_T^M(\theta)\right]$ to obtain

$$\gamma_T^{\theta_k}(\theta) = -\frac{1}{2}\text{trace}\left[\widehat{\Gamma}_T(\theta)^{-1}\frac{\partial}{\partial\theta_k}\widehat{V}_T(\theta)\right] + \frac{1}{2}\text{trace}\left[\widehat{\Gamma}_T(\theta)^{-1}\frac{\partial}{\partial\theta_k}\widehat{\Gamma}_T(\theta)\widehat{\Gamma}_T(\theta)^{-1}\widehat{V}_T(\theta)\right] \tag{C.10}$$

where

$$\begin{aligned}\widehat{\Gamma}_T^{\theta_k}(\theta) \equiv \frac{\partial}{\partial\theta_k}\widehat{\Gamma}_T(\theta) = &-\widehat{H}_T^{\theta_k}(\theta)k_T + A_T^{\theta_k}(\theta)(1-k_T) \\ &- \left(\widehat{H}_T(\theta) + A_T(\theta)\right)\left(\frac{\partial k_T}{\partial(-\widehat{\lambda}_{\max})}\frac{\partial(-\widehat{\lambda}_{\max})}{\partial\theta_k} + \frac{\partial k_T}{\partial(-\widehat{\lambda}_{\min})}\frac{\partial(-\widehat{\lambda}_{\min})}{\partial\theta_k}\right).\end{aligned} \tag{C.11}$$

We then need that $\gamma_T^{\theta_k}(\theta_0) = \gamma_T^{\theta_k} \xrightarrow{p} \bar{\gamma}_T^{\theta_k}$ for some function $\bar{\gamma}_T^{\theta_k}$.

Under conditions (i)-(iii), (v), and (vi) of Assumption 1, $\widehat{H}_T \xrightarrow{p} H_0$. It follows by the continuous mapping theorem that $\widehat{\lambda}_{\min} \xrightarrow{p} \lambda_{\min}$ and $\widehat{\lambda}_{\max} \xrightarrow{p} \lambda_{\max}$. Since $\delta_T \to 0$, $h_T \to 0$, and $\Delta_T \to \infty$, $k_T \to 1$, $\frac{\partial k_T}{\partial(-\widehat{\lambda}_{\min})} \to 0$, and $\frac{\partial k_T}{\partial(-\widehat{\lambda}_{\max})} \to 0$. In addition, $\frac{\partial\lambda_{\min}}{\partial\theta_k} \xrightarrow{p} m_1$ and $\frac{\partial\lambda_{\max}}{\partial\theta_k} \xrightarrow{p} m_2$ for some numbers $m_1$ and $m_2$ since $\lambda_{\min}$ and $\lambda_{\max}$ are analytic functions of the elements of $\widehat{H}_T$ under condition (iv) of Assumption 3. We may then conclude that

$$\widehat{\Gamma}_T^{-1} \xrightarrow{p} -H_0^{-1} \quad \text{and} \quad \widehat{\Gamma}_T^{\theta_k} \xrightarrow{p} -H_0^{\theta_k}. \tag{C.12}$$

Next note that

$$\frac{\partial}{\partial\theta_k}\widehat{V}_T(\theta) = \frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\left(\ell_t^{\theta\theta_k}(\theta)(\ell_{t-j}^{\theta}(\theta))' + \ell_t^{\theta}(\theta)(\ell_{t-j}^{\theta\theta_k}(\theta))'\right).$$

Plugging this expression into $\frac{1}{2}\text{trace}\left[\widehat{\Gamma}_T(\theta)^{-1}\frac{\partial}{\partial\theta_k}\widehat{V}_T(\theta)\right]$ and rearranging terms via cyclic permutations under the trace then yields

$$\frac{1}{2}\text{trace}\left[\widehat{\Gamma}_T(\theta)^{-1}\frac{\partial}{\partial\theta_k}\widehat{V}_T(\theta)\right] = \text{trace}\left[\widehat{\Gamma}_T(\theta)^{-1}\left(\frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^{\theta}(\theta)(\ell_{t-j}^{\theta\theta_k}(\theta))'\right)\right].$$

Then, evaluating at $\theta_0$ and adding and subtracting terms, we have

$$\begin{aligned}
&\frac{1}{2}\text{trace}\left[\widehat{\Gamma}_T^{-1}\frac{\partial}{\partial\theta_k}\widehat{V}_T\right] \\
&\qquad = \text{trace}\left[\widehat{\Gamma}_T^{-1}\left(\frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^\theta(\ell_{t-j}^{\theta\theta_k}-\mathrm{E}[\ell_{t-j}^{\theta\theta_k}])'\right)\right] \qquad \text{(C.13)} \\
&\qquad\qquad + \text{trace}\left[\widehat{\Gamma}_T^{-1}\left(\frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^\theta(\mathrm{E}[\ell_{t-j}^{\theta\theta_k}])'\right)\right].
\end{aligned}$$

Next,

$$\begin{aligned}
\mathrm{E}\|\frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^\theta(\mathrm{E}[\ell_{t-j}^{\theta\theta_k}])'\| &\le \frac{1}{T}\sum_{j=-m}^{m}\mathrm{E}\|\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^\theta(\mathrm{E}[\ell_{t-j}^{\theta\theta_k}])'\| \\
&\le \frac{1}{T}\sum_{j=-m}^{m}\left(\mathrm{E}\|\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^\theta(\mathrm{E}[\ell_{t-j}^{\theta\theta_k}])'\|^2\right)^{1/2} \\
&\le \frac{CT^{1/2}(2m+1)}{T} \to 0 \quad \text{if} \quad m = o(T^{1/2}) \qquad \text{(C.14)}
\end{aligned}$$

for some constant $C$ where the inequality in (C.14) follows under the mixing and moment conditions in Assumption 3 using, for example, Doukhan (1994) Theorem 2. Also, under the assumed mixing and moment conditions, we have

$$\widehat{V}_T \xrightarrow{p} -H_0 \qquad \text{(C.15)}$$

and

$$\frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_t^\theta(\ell_{t-j}^{\theta\theta_k}-\mathrm{E}[\ell_{t-j}^{\theta\theta_k}])' \xrightarrow{p} \lim_{T\to\infty}\mathrm{E}\left[T^{-1}\sum_{t=1}^{T}\ell_t^\theta\sum_{t=1}^{T}\ell_t^{\theta\theta_k}\right]. \qquad \text{(C.16)}$$

from, for example, the modification of Andrews (1991) Theorem 1(a) for nonstationary data discussed in Andrews (1991) Section 8.

Using (C.12), (C.13), (C.14), and (C.16) gives

$$\frac{1}{2}\text{trace}\left[\widehat{\Gamma}_T(\theta)^{-1}\frac{\partial}{\partial\theta_k}\widehat{V}_T(\theta)\right] \xrightarrow{p} \text{trace}\left[-H_0^{-1}\lim_{T\to\infty}\mathrm{E}\left(T^{-1}\sum_{t=1}^{T}\ell_t^\theta\sum_{t=1}^{T}\ell_t^{\theta\theta_k}\right)\right], \qquad \text{(C.17)}$$

and using (C.12) and (C.15) yields

$$\frac{1}{2}\text{trace}\left[\widehat{\Gamma}_T(\theta)^{-1}\frac{\partial}{\partial\theta_k}\widehat{\Gamma}_T(\theta)\widehat{\Gamma}_T(\theta)^{-1}\widehat{V}_T(\theta)\right] \xrightarrow{p} \frac{1}{2}\text{trace}\left[H_0^{-1}H_0^{\theta_k}\right]. \qquad \text{(C.18)}$$

We then have

$$\gamma_T^{\theta_k} \xrightarrow{p} \bar{\gamma}_T^{\theta_k} = -\text{trace}\left[-H_0^{-1} \lim_{T\to\infty} \mathrm{E}\left(T^{-1}\sum_{t=1}^{T}\ell_t^{\theta}\sum_{t=1}^{T}\ell_t^{\theta\theta_k}\right)\right] + \frac{1}{2}\text{trace}\left[H_0^{-1}H_0^{\theta_k}\right].$$

We also have by stacking the above derivatives that

$$\frac{\partial}{\partial\theta}\log[\widetilde{\pi}_T^M(\theta)] - \left(-\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] + \frac{1}{2}\sum_{j=1}^{k}\text{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]e_j\right) = o_p(1).$$

It remains to be shown that the conditions of Assumption 3 are sufficient for the derivative and moment conditions in condition (viii) of Assumption 1. That the log-prior is six times differentiable follows from the definition of $\widehat{\Gamma}_T(\theta)$ and the eighth-order differentiability of the likelihood assumed in condition (vi) of Assumption 3. Verifying that for $0 \leq j \leq 6$ $\sup_{\theta\in\mathcal{G}}\mathrm{E}\|\Delta^j\gamma_T(\theta)\| = o(T^{1/2})$ is a tedious but straightforward exercise. We verify that $\mathrm{E}\|\Delta^j\gamma_T(\theta)\| = o(T^{1/2})$ for $j = 0$ and $j = 1$. The result for the higher-order derivatives follows in a similar fashion and we omit it for brevity.

For any $\theta \in \mathcal{G}$,

$$\begin{aligned}
\mathrm{E}\|\text{trace}\left[\widehat{\Gamma}_T^{-1}(\theta)\widehat{V}_T(\theta)\right]\| &\leq \left(\frac{p}{\delta_T^2}\right)^{1/2}\frac{1}{T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}\left(\mathrm{E}\|\ell_t^{\theta}\|^2\mathrm{E}\|\ell_{t-j}^{\theta}\|^2\right)^{1/2} \\
&\leq \frac{C}{\delta_T}(2m - m^2/T - m/T + 1) \\
&= o(T^{1/2})
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz and Triangle inequalities, $|K(x)| \leq 1$, and $\|\widehat{\Gamma}_T^{-1}\| \leq \left(\frac{p}{\delta_T^2}\right)^{1/2}$; the second inequality follows from condition (vii) of Assumption 3; and the final equality follows from the rate conditions imposed in condition (iii) of Assumption 3. It is then immediate that $\sup_{\theta\in\mathcal{G}}\mathrm{E}\|\gamma_T(\theta)\| = o(T^{1/2})$.

The vector of first partial derivatives of $\gamma_T(\theta)$ may be written as

$$\gamma_T^{\theta}(\theta) = -\frac{1}{2T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}K(j/m)\ell_{t-j}^{\theta\theta}(\theta)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta) \tag{C.19}$$

$$-\frac{1}{2T}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}K(j/m)\ell_t^{\theta\theta}(\theta)\widehat{\Gamma}_T(\theta)^{-1}\ell_{t-j}^{\theta}(\theta) \tag{C.20}$$

$$+\frac{1}{2T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}\widehat{\Gamma}_T^{\theta_k}(\theta)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k. \tag{C.21}$$

That the expectations of the norm of the terms given in (C.19) and (C.20) are $o(T^{1/2})$ follow from an argument identical to that used above to show that $\mathrm{E}\|\gamma_T(\theta)\| = o(T^{1/2})$.

We also have

$$\begin{aligned}
&\mathrm{E}\|\frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}\widehat{H}_T^{\theta_k}(\theta)k_T\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k\| \\
&\quad\leq \frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}\left(\mathrm{E}\|\ell_{t-j}^{\theta}(\theta)\|^3\mathrm{E}\|\widehat{H}_T^{\theta_k}(\theta)\|^3\mathrm{E}\|\ell_t^{\theta}(\theta)\|^3\right)^{1/3}\frac{C}{\delta_T^2} \\
&\quad\leq \frac{C}{\delta_T^2}(2m - m^2/T - m/T + 1) = o(T^{1/2}) \qquad \text{(C.22)}
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz and Triangle inequalities, $|K(x)| \leq 1$, $|k_T| \leq 1$, and $\|\widehat{\Gamma}_T^{-1}\| \leq \left(\frac{p}{\delta_T^2}\right)^{1/2}$; the second inequality from condition (vii) of Assumption 3; and the final equality from the rate conditions imposed in condition (iii) of Assumption 3. Similarly, using Assumption 3.(i),

$$\begin{aligned}
&\mathrm{E}\|\frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}A_T^{\theta_k}(\theta)(1-k_T)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k\| \\
&\quad\leq \frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}\left(\mathrm{E}\|\ell_{t-j}^{\theta}(\theta)\|^2\mathrm{E}\|\ell_t^{\theta}(\theta)\|^2\right)^{1/2}\frac{C}{\delta_T^2} \\
&\quad\leq \frac{C}{\delta_T^2}(2m - m^2/T - m/T + 1) = o(T^{1/2}). \qquad \text{(C.23)}
\end{aligned}$$

Assumption 3.(iv) implies that $\lambda_{\min}$ and $\lambda_{\max}$ are analytic functions of the elements of $\widehat{H}_T(\theta)$, so it follows under Assumption 3.(vii) that $\mathrm{E}\|\frac{\partial(-\widehat{\lambda}_{\max})}{\partial\theta_k}\| \leq \Delta$ and $\|\frac{\partial(-\widehat{\lambda}_{\min})}{\partial\theta_k}\| \leq \Delta$. Then, as above

$$\begin{aligned}
&\mathrm{E}\|\frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}} K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}\widehat{H}_T^{\theta_k}(\theta)\left(\frac{\partial k_T}{\partial(-\widehat{\lambda}_{\max})}\frac{\partial(-\widehat{\lambda}_{\max})}{\partial\theta_k}\right)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k\| \\
&\quad\leq \frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}\left(\mathrm{E}\|\ell_{t-j}^{\theta}(\theta)\|^4\mathrm{E}\|\widehat{H}_T^{\theta_k}(\theta)\|^4\mathrm{E}\|\frac{\partial(-\widehat{\lambda}_{\max})}{\partial\theta_k}\|^4\mathrm{E}\|\ell_t^{\theta}(\theta)\|^4\right)^{1/4}\frac{C}{\delta_T^2} \\
&\quad\leq \frac{C}{\delta_T^2}(2m - m^2/T - m/T + 1) = o(T^{1/2}), \qquad \text{(C.24)}
\end{aligned}$$

$$\mathrm{E}\|\frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}\widehat{H}_T^{\theta_k}(\theta)\left(\frac{\partial k_T}{\partial(-\widehat{\lambda}_{\min})}\frac{\partial(-\widehat{\lambda}_{\min})}{\partial\theta_k}\right)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k\|$$

$$\leq \frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}\left(\mathrm{E}\|\ell_{t-j}^{\theta}(\theta)\|^4\mathrm{E}\|\widehat{H}_T^{\theta_k}(\theta)\|^4\mathrm{E}\|\frac{\partial(-\widehat{\lambda}_{\min})}{\partial\theta_k}\|^4\mathrm{E}\|\ell_t^{\theta}(\theta)\|^4\right)^{1/4}\frac{C}{\delta_T^2}$$

$$\leq \frac{C}{\delta_T^2}(2m-m^2/T-m/T+1)=o(T^{1/2}), \tag{C.25}$$

$$\mathrm{E}\|\frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}A_T^{\theta_k}(\theta)\left(\frac{\partial k_T}{\partial(-\widehat{\lambda}_{\max})}\frac{\partial(-\widehat{\lambda}_{\max})}{\partial\theta_k}\right)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k\|$$

$$\leq \frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}\left(\mathrm{E}\|\ell_{t-j}^{\theta}(\theta)\|^3\mathrm{E}\|\frac{\partial(-\widehat{\lambda}_{\max})}{\partial\theta_k}\|^3\mathrm{E}\|\ell_t^{\theta}(\theta)\|^3\right)^{1/3}\frac{C}{\delta_T^2}$$

$$\leq \frac{C}{\delta_T^2}(2m-m^2/T-m/T+1)=o(T^{1/2}), \tag{C.26}$$

and

$$\mathrm{E}\|\frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}A_T^{\theta_k}(\theta)\left(\frac{\partial k_T}{\partial(-\widehat{\lambda}_{\min})}\frac{\partial(-\widehat{\lambda}_{\min})}{\partial\theta_k}\right)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k\|$$

$$\leq \frac{1}{T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}\left(\mathrm{E}\|\ell_{t-j}^{\theta}(\theta)\|^3\mathrm{E}\|\frac{\partial(-\widehat{\lambda}_{\min})}{\partial\theta_k}\|^3\mathrm{E}\|\ell_t^{\theta}(\theta)\|^3\right)^{1/3}\frac{C}{\delta_T^2}$$

$$\leq \frac{C}{\delta_T^2}(2m-m^2/T-m/T+1)=o(T^{1/2}). \tag{C.27}$$

Then plugging the expression for $\widehat{\Gamma}_T^{\theta_k}(\theta)$ given in (C.11) into the term given in (C.21) and using (C.22)-(C.27) gives that for all $\theta\in\mathcal{G}$,

$$\mathrm{E}\|\frac{1}{2T}\sum_{k=1}^{p}\sum_{j=-m}^{m}\sum_{t=\max\{1,j+1\}}^{\max\{T,T+j\}}K(j/m)\ell_{t-j}^{\theta}(\theta)'\widehat{\Gamma}_T(\theta)^{-1}\widehat{\Gamma}_T^{\theta_k}(\theta)\widehat{\Gamma}_T(\theta)^{-1}\ell_t^{\theta}(\theta)e_k\|=o(T^{1/2}).$$

It then follows that $\sup_{\theta\in\mathcal{G}}\mathrm{E}\|\gamma_T^{\theta}(\theta)\|=o(T^{1/2})$. As noted above, results for the second through sixth derivatives are obtained in a similar fashion. ■

## Appendix D. Derivation of AR(1) Priors

In the following, we outline the derivation of the non-data-dependent priors for the dynamic panel model:

$$y_{it} = \rho y_{it-1} + \alpha_i + u_{it}$$

where $u_{it} \sim N(0, \sigma^2)$ are iid for all $i$ and $t$, $|\rho| < 1$, $i = 1, ..., N$, and $t = 1, ..., T$. In the derivation, we do not make use of the panel data simplification discussed in Section 4.3 as derivation of the priors making use of the simplification is discussed in Bester and Hansen (2005) and is similar to the derivation below. In deriving the prior, we treat $N$ as fixed and consider asymptotics as $T \to \infty$. Thus, by setting $N = 1$, the prior applies to the simple univariate autoregression as well.

We begin by noting that

$$E[y_{it}|\alpha_i] = \alpha_i/(1-\rho) \tag{D.1}$$

and

$$E[y_{it}^2|\alpha_i] = (\alpha_i/(1-\rho))^2 + \sigma^2/(1-\rho^2). \tag{D.2}$$

We also have that the log-likelihood for a single observation is given by

$$\ell_{it}(\theta) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}u_{it}^2 \tag{D.3}$$

where $u_{it} = y_{it} - \alpha_i - \rho y_{it-1}$. Deriving the differential equations that define the priors then requires computing $\mathrm{E}[\sqrt{T}\widehat{H}_T\psi_1]$ and $\operatorname{trace}\left[H_T^{\theta_j}(H_T)^{-1}\right]$ for $j = 1, 2, ..., N+2$ with $\theta = (\sigma^2, \rho, \alpha_1, ...\alpha_N)'$.

Differentiating (D.3) with respect to $\theta$ yields

$$\ell_{it}^{\theta}(\theta) = \begin{pmatrix} -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}u_{it}^2 \\ \frac{1}{\sigma^2}y_{it-1}u_{it} \\ \frac{1}{\sigma^2}u_{it}e_i \end{pmatrix} \tag{D.4}$$

where $e_i$ is the i$^{\text{th}}$ unit vector, and differentiating again gives the per observation hessian

$$\ell_{it}^{\theta\theta}(\theta) = \begin{pmatrix} \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}u_{it}^2 & -\frac{1}{\sigma^4}y_{it-1}u_{it} & -\frac{1}{\sigma^4}u_{it}e_i' \\ -\frac{1}{\sigma^4}y_{it-1}u_{it} & -\frac{1}{\sigma^2}y_{it-1}^2 & -\frac{1}{\sigma^2}y_{it-1}e_i' \\ -\frac{1}{\sigma^4}u_{it}e_i & -\frac{1}{\sigma^2}y_{it-1}e_i & -\frac{1}{\sigma^2}e_ie_i' \end{pmatrix}. \tag{D.5}$$

It then follows that

$$-H_T^{-1} = \begin{pmatrix} \frac{2\sigma^4}{N} & 0 & 0_{1\times N} \\ 0 & \frac{1-\rho^2}{N} & -\frac{1-\rho^2}{N}\sum_{i=1}^N \frac{\alpha_i}{1-\rho}e_i' \\ 0_{N\times 1} & -\frac{1-\rho^2}{N}\sum_{i=1}^N \frac{\alpha_i}{1-\rho}e_i & (1-\rho^2)\left[\frac{\sigma^2}{1-\rho^2}I_N + \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N \frac{\alpha_i\alpha_j}{(1-\rho)^2}e_ie_j'\right] \end{pmatrix}. \tag{D.6}$$

Differentiating (D.5) again with respect to $\sigma^2$, summing, and taking expectations gives

$$H_T^{\sigma^2} = \begin{pmatrix} \frac{2N}{\sigma^6} & 0 & 0_{1\times N} \\ 0 & \frac{1}{\sigma^4}\left[\sum_{i=1}^N \left(\frac{\alpha_i}{1-\rho}\right)^2 + \frac{N\sigma^2}{1-\rho^2}\right] & \frac{1}{\sigma^4}\sum_{i=1}^N \frac{\alpha_i}{1-\rho} e_i' \\ 0_{N\times 1} & \frac{1}{\sigma^4}\sum_{i=1}^N \frac{\alpha_i}{1-\rho} e_i & \frac{1}{\sigma^4} I_N \end{pmatrix}. \tag{D.7}$$

It then follows by multiplying (D.6) and (D.7) and some algebra that

$$\text{trace}(-H_T^{-1} H_T^{\sigma^2}) = \frac{5+N}{\sigma^2}. \tag{D.8}$$

Similarly, we have that

$$H_T^{\rho} = \begin{pmatrix} 0 & \frac{1}{\sigma^4}\left[\sum_{i=1}^N \left(\frac{\alpha_i}{1-\rho}\right)^2 + \frac{N\sigma^2}{1-\rho^2}\right] & \frac{1}{\sigma^4}\sum_{i=1}^N \frac{\alpha_i}{1-\rho} e_i' \\ \frac{1}{\sigma^4}\left[\sum_{i=1}^N \left(\frac{\alpha_i}{1-\rho}\right)^2 + \frac{N\sigma^2}{1-\rho^2}\right] & 0 & 0_{1\times N} \\ \frac{1}{\sigma^4}\sum_{i=1}^N \frac{\alpha_i}{1-\rho} e_i & 0_{N\times 1} & 0_{N\times N} \end{pmatrix} \tag{D.9}$$

and that, for each $i = 1, ..., N$,

$$H_T^{\alpha_i} = \begin{pmatrix} 0 & \frac{1}{\sigma^4}\frac{\alpha_i}{1-\rho} & \frac{1}{\sigma^4} e_i' \\ \frac{1}{\sigma^4}\frac{\alpha_i}{1-\rho} & 0 & 0_{1\times N} \\ \frac{1}{\sigma^4} e_i & 0_{N\times 1} & 0_{N\times N} \end{pmatrix} \tag{D.10}$$

from which

$$\text{trace}(-H_T^{-1} H_T^{\rho}) = 0 \tag{D.11}$$

and

$$\text{trace}(-H_T^{-1} H_T^{\alpha_i}) = 0 \quad \forall\ i \tag{D.12}$$

follow.

Computation of the prior then requires that we compute

$$\text{E}\left[\sqrt{T}\widehat{H}_T \psi_1\right] = \sum_{i=1}^N \text{E}\left[T^{-1}\sum_{t=1}^T\sum_{\tau=1}^T \ell_{it}^{\theta\theta}\psi_{i\tau}\right]$$

where the equality results from assuming independence across $i$ and $\psi_{it}$ is the per observation influence function obtained by premultiplying (D.4) by (D.6):

$$\psi_{it} = \begin{pmatrix} \frac{1}{N}\left(u_{it}^2 - \sigma^2\right) \\ \frac{1-\rho^2}{N\sigma^2} y_{it-1}u_{it} - \frac{1-\rho^2}{N\sigma^2}\frac{\alpha_i}{1-\rho}u_{it} \\ -\frac{1-\rho^2}{N\sigma^2} y_{it-1}u_{it}\sum_{j=1}^N \frac{\alpha_j}{1-\rho} e_j + \frac{1-\rho^2}{N\sigma^2}\left[\frac{N\sigma^2}{1-\rho^2} + \left(\frac{\alpha_i}{1-\rho}\right)^2\right] u_{it} e_i \end{pmatrix}. \tag{D.13}$$

Using the equation for $\ell_{it}^{\theta\theta}$ in (D.5) above and equation (D.13) it follows from a straightforward but rather tedious calculation that

$$\mathrm{E}\left[\sqrt{T}\widehat{H}_T\psi_1\right] = \begin{pmatrix} \frac{3+N}{\sigma^2} \\ \frac{1}{T}\sum_{t=1}^{T-1}(T-t)\left(N\rho^{t-1}+\rho^{2t-1}\right) \\ 0_{N\times 1} \end{pmatrix}. \tag{D.14}$$

The priors may then be obtained by plugging (D.8), (D.11), (D.12), and (D.14) into the differential equations (4.1) and (4.2) and integrating. Performing this calculation yields the bias-reducing prior for the joint mode

$$\pi^M(\theta) = (\sigma^2)^{\frac{N+1}{2}} \exp\left[\frac{N}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho^t\right] \exp\left[\frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{2t}\rho^{2t}\right] \tag{D.15}$$

and for the posterior mean

$$\pi^E(\theta) = (\sigma^2)^{-2} \exp\left[\frac{N}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho^t\right] \exp\left[\frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{2t}\rho^{2t}\right]. \tag{D.16}$$

These priors have some noteworthy features. When $|\rho| < 1$, somewhat simplified expressions may be obtained by noting that

$$\pi^M(\theta) \to \frac{(\sigma^2)^{\frac{N+1}{2}}}{(1-\rho)^N(1-\rho^2)^{1/2}}$$

and

$$\pi^E(\theta) \to \frac{(\sigma^2)^{-2}}{(1-\rho)^N(1-\rho^2)^{1/2}}$$

as $T \to \infty$. While these priors may be somewhat simpler to work with, it should be noted that the poles at 1 and -1 may result in undesirable performance of the resulting estimators when $T$ is small or moderate. Also, while the priors were derived under the assumption that $|\rho| < 1$, their use is not necessarily restricted to that case. Indeed, they bear a close resemblence to the orthogonalization suggested by Lancaster (2002) for the nonstationary panel case which is equivalent to using the prior

$$\pi(\theta) = \exp\left[\frac{N}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho^t\right]$$

and then integrating out over the fixed effects. As $N \to \infty$, this prior is equivalent to $\pi^E(\theta)$, and as noted by Lancaster (2002), produces consistent estimates of the autoregressive coefficient even when $T$ is fixed.

It is also important to note that unless $|\rho| < 1$, the prior is explosive. This behavior means that some sort of truncation of the prior is necessary for Condition (viii) of Assumption 1 to be satisfied if one is not willing to impose $|\rho| < 1$ a priori. To illustrate, let

$$b_1(\rho) = \frac{N}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\rho^t$$

and

$$b_2(\rho) = \frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{2t}\rho^{2t}.$$

It is obvious that $b_1(\rho)$ and $b_2(\rho)$ diverge for $|\rho| \geq 1$ and that $b_1(1)/\sqrt{T} \to 0$ and $b_2(1)/\sqrt{T} \to 0$ as $T \to \infty$. Thus, if $b_1(\rho)$ and $b_2(\rho)$ are truncated at $b_1(1)$ and $b_2(1)$, the condition is satisfied. There are a number of ways to achieve this. A natural way is to replace $b_1(\rho)$ and $b_2(\rho)$ in the priors with

$$\widetilde{b}_1(\rho) = [b_1(\rho)k(\rho) + b_1(1)(1-k(\rho))]1(\rho < 1) + [b_1(2-\rho)k(2-\rho) + b_1(1)(1-k(2-\rho))]1(\rho \geq 1)$$

and

$$\widetilde{b}_2(\rho) = [b_2(\rho)k(\rho) + b_2(1)(1-k(\rho))]1(\rho < 1) + [b_2(2-\rho)k(2-\rho) + b_2(1)(1-k(2-\rho))]1(\rho \geq 1)$$

where $k(\rho)$ is a 6 times continuously differentiable weighting function with $k(x) = 0$ for $x \geq 1$ and $k(x) = 1$ for $x < 1 - h_T$ where $h_T \to 0$ as $T \to \infty$; for example, one could use the weighting function in Appendix A. The priors formed with these substitutions are appealing in that they are proper and will reduce bias in the posterior mean or mode for $\rho$ as long as $|\rho| < 1$. For $|\rho| \geq 1$, they will result in a bias toward stationarity, though this should be slight as the likelihood concentrates very rapidly when $\rho \geq 1$. Alternatively, one could use

$$\widetilde{b}_1(\rho) = [b_1(\rho)k(\rho) + b_1(1)(1-k(\rho))]1(\rho < 1) + b_1(1)1(\rho \geq 1)$$

and

$$\widetilde{b}_2(\rho) = [b_2(\rho)k(\rho) + b_2(1)(1-k(\rho))]1(\rho < 1) + b_2(1)1(\rho \geq 1).$$

In this case, the resulting priors are not proper but will not induce downward bias when $\rho \geq 1$. We considered both versions of the prior in preliminary simulations. For models where the true value of $\rho$ was less than 1, both forms of the truncated prior produced nearly identical results. As such, we chose to report only results based on the proper version in the paper.

The truncation discussed above seems to be fairly reasonable in this context. The expansions derived in Section 3 of this paper are only valid if the model is non-explosive, and for $\rho \geq 1$, the usual estimates of $\rho$ are superconsistent. That is, for $\rho \geq 1$, the likelihood concentrates rapidly at the true parameter value and bias is not a major concern, though the resulting distributions are nonstandard and inference remains problematic. It may be interesting to consider extensions of the main results of the paper to nonstationary cases. As noted above, the likelihood will concentrate

very rapidly in these models, and to remove bias, the prior will need to diverge more rapidly than $\sqrt{T}$ as well. In this sense, it seems likely the priors given in (D.15) and (D.16) may remove the higher-order bias in these settings under a much looser constraint on the truncation than the one considered above. While potentially interesting, extending our results in this direction is beyond the scope of the present paper.

## References


Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

Arellano, M. (2003): *Panel Data Econometrics*. Oxford University Press.

Arellano, M., and J. Hahn (2005): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," Invited Lecture, Econometric Society World Congress, London.

Bartlett, M. S. (1953): "Approximate Confidence Intervals II," *Biometrika*, 40, 306–317.

Bester, A. C., and C. Hansen (2005): "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects," Mimeo.

Cox, D. R., and E. J. Snell (1968): "A general definition of residuals (with discussion)," *Journal of the Royal Statistical Society*, 30, 248–275.

Doukhan, P. (1994): *Mixing: Properties and Examples*, vol. 85 of *Lecture Notes in Statistics (Springer-Verlag)*. New York: Springer-Verlag.

Firth, D. (1993): "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80(1), 27–38.

Ghosh, J. K., and R. Mukerjee (1992): "Non-Informative Priors," in *Bayesian Statistics 4*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 195–210. Clarendon Press: Oxford, UK.

Hahn, J., and G. Kuersteiner (2004): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," Mimeo.

Hahn, J., and G. M. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both N and T Are Large," *Econometrica*, 70(4), 1639–1657.

Hahn, J., and W. K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Jeffreys, H. (1946): "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London, Series A*, 186, 453–461.

Kass, R. E., L. Tierney, and J. B. Kadane (1990): "The Validity of Posterior Expansions Based on Laplaces's Method," in *Bayesian and Likelihood Methods in Statistics and Econometrics*, ed. by S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner. Elsevier Science Publishers B.V.: North-Holland.

Kass, R. E., and L. Wasserman (1996): "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91(435), 1343–1370.

Lancaster, T. (2002): "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647–666.

Mukerjee, R., and D. K. Dey (1993): "Frequentist Validity of Posterior Quantiles in the Presence of a Nuisance Parameter: Higher-Order Asymptotics," *Biometrika*, 80, 499–505.

Newey, W. K., and R. J. Smith (2001): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," Mimeo.

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

Neyman, J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1–32.

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49(6), 1417–1426.

Phillips, P. C. B. (1991): "Bayesian Routes and Unit Roots: *de rebus prioribus semper est disputandum*," *Journal of Applied Econometrics*, 6, 435–474.

Reid, N., R. Mukerjee, and D. A. S. Fraser (2002): "Some Aspects of Matching Priors," in *Mathematical Statistics and Applications: Festschrift for C. VanEeden*, ed. by M. Moore, S. Froda, and C. Léger. Lecture Notes Monograph Series 42, Institute of Mathematical Statistics, Hayward.

Rilstone, P., V. K. Srivastava, and A. Ullah (1996): "The Second-Order Bias and Mean Squared Error of Nonlinear Estimators," *Journal of Econometrics*, 75, 369–395.
Shaman, P., and R. A. Stine (1988): "The Bias of Autoregressive Coefficient Estimators," *Journal of the American Statistical Association*, 83, 842–848.
Shenton, L. R., and P. A. Wallington (1962): "The bias of moment estimators with an application to the negative binomial distribution," *Biometrika*, 49, 193–204.
Sweeting, T. J. (2004): "On the Implementation of Local Probability Matching Priors for Interest Parameters," Research Report No. 241, Department of Statistical Science, University College London.
Wasserman, L. (2000): "Asymptotic Inference for Mixture Models Using Data-Dependent Priors," *Journal of the Royal Statistical Society, Series B*, 62(1), 159–180.
Welch, B., and H. W. Peers (1963): "On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods," *Journal of the Royal Statistical Society, Series B*, 25, 318–329.
White, H. (2001): *Asymptotic Theory for Econometricians*. San Diego: Academic Press, revised edn.
Wooldridge, J. M. (1994): "Estimation and Inference for Dependent Processes," in *Handbook of Econometrics. Volume 4*, ed. by R. F. Engle, and D. L. McFadden. Elsevier: North-Holland.

TABLE 1. Monte Carlo Results from Simulated AR(1) Model ($N = 1$)

| | T = 50 | | | | T = 200 | | | |
|---|---|---|---|---|---|---|---|---|
| Estimator | Bias | RMSE | MAD | RP(.05) | Bias | RMSE | MAD | RP(.05) |
| | | | | $\rho = 0.6$ | | | | |
| $\widehat{\rho}^{M}$ | -0.054 | 0.177 | 0.100 | 0.942 | -0.012 | 0.083 | 0.047 | 0.947 |
| $\widehat{\rho}^{M}_{DD}$ | -0.028 | 0.170 | 0.095 | 0.956 | -0.002 | 0.083 | 0.047 | 0.951 |
| $\widehat{\rho}^{M}_{CF}$ | -0.004 | 0.183 | 0.102 | 0.932 | -0.001 | 0.083 | 0.047 | 0.950 |
| $\widehat{\rho}^{E}$ | -0.054 | 0.177 | 0.101 | 0.958 | -0.013 | 0.084 | 0.047 | 0.951 |
| $\widehat{\rho}^{E}_{DD}$ | -0.027 | 0.170 | 0.095 | 0.940 | -0.002 | 0.083 | 0.047 | 0.941 |
| $\widehat{\rho}^{E}_{CF}$ | 0.011 | 0.192 | 0.110 | 0.949 | -0.001 | 0.084 | 0.047 | 0.952 |
| $\widehat{\rho}^{M}_{BCDD}$ | -0.021 | 0.170 | 0.095 | 0.952 | -0.003 | 0.082 | 0.046 | 0.952 |
| $\widehat{\rho}^{M}_{BCCF}$ | -0.041 | 0.183 | 0.103 | 0.937 | -0.009 | 0.084 | 0.047 | 0.944 |
| | | | | $\rho = 0.9$ | | | | |
| $\widehat{\rho}^{M}$ | -0.081 | 0.155 | 0.092 | 0.863 | -0.020 | 0.054 | 0.030 | 0.933 |
| $\widehat{\rho}^{M}_{DD}$ | -0.066 | 0.141 | 0.080 | 0.920 | -0.014 | 0.050 | 0.027 | 0.951 |
| $\widehat{\rho}^{M}_{CF}$ | 0.009 | 0.155 | 0.095 | 0.942 | -0.004 | 0.052 | 0.028 | 0.934 |
| $\widehat{\rho}^{E}$ | -0.081 | 0.155 | 0.092 | 0.932 | -0.020 | 0.054 | 0.030 | 0.941 |
| $\widehat{\rho}^{E}_{DD}$ | -0.066 | 0.142 | 0.080 | 0.885 | -0.014 | 0.050 | 0.027 | 0.938 |
| $\widehat{\rho}^{E}_{CF}$ | -0.013 | 0.120 | 0.061 | 0.972 | 0.001 | 0.054 | 0.030 | 0.950 |
| $\widehat{\rho}^{M}_{BCDD}$ | -0.062 | 0.138 | 0.077 | 0.921 | -0.015 | 0.051 | 0.028 | 0.949 |
| $\widehat{\rho}^{M}_{BCCF}$ | -0.052 | 0.150 | 0.081 | 0.893 | -0.012 | 0.052 | 0.028 | 0.943 |

Monte Carlo results for univariate AR(1) simulation. Results are for the autoregressive parameter where the true value is .6 in the upper panel and .9 in the lower panel. Bias, root mean squared error (RMSE), mean absolute deviation (MAD), and coverage probabilities of 95% level interval estimates (CP(.95)) are reported. $\widehat{\rho}^{M}_{DD}$, $\widehat{\rho}^{M}_{CF}$, $\widehat{\rho}^{E}_{DD}$, and $\widehat{\rho}^{E}_{CF}$ are Bayesian estimators with bias-reducing priors. $\widehat{\rho}^{M}$ is the MLE, and $\widehat{\rho}^{E}$ is the posterior mean under a flat prior. The remaining estimators are bias-corrected maximum likelihood estimators. The number of simulations is 1000.

TABLE 2. Monte Carlo Results from Simulated Dynamic Panel Model ($N = 100$)

| | T = 4 | | | | T = 12 | | | |
|---|---|---|---|---|---|---|---|---|
| Estimator | Bias | RMSE | MAD | CP(.95) | Bias | RMSE | MAD | CP(.95) |
| | | | | $\rho = 0.6$ | | | | |
| $\widehat{\rho}^{M}$ | -0.448 | 0.454 | 0.448 | 0.000 | -0.148 | 0.154 | 0.148 | 0.000 |
| $\widehat{\rho}^{M}_{DD}$ | -0.366 | 0.374 | 0.366 | 0.000 | -0.102 | 0.108 | 0.102 | 0.031 |
| $\widehat{\rho}^{M}_{CF}$ | 0.059 | 0.272 | 0.142 | 0.617 | 0.000 | 0.049 | 0.027 | 0.877 |
| $\widehat{\rho}^{E}$ | -0.448 | 0.454 | 0.448 | 0.000 | -0.148 | 0.154 | 0.148 | 0.000 |
| $\widehat{\rho}^{E}_{DD}$ | -0.366 | 0.374 | 0.366 | 0.000 | -0.102 | 0.109 | 0.102 | 0.019 |
| $\widehat{\rho}^{E}_{CF}$ | 0.050 | 0.250 | 0.134 | 0.575 | 0.000 | 0.049 | 0.027 | 0.820 |
| $\widehat{\rho}^{M}_{BCDD}$ | -0.206 | 0.260 | 0.208 | 0.169 | -0.119 | 0.126 | 0.119 | 0.011 |
| $\widehat{\rho}^{M}_{BCCF}$ | -0.160 | 0.186 | 0.160 | 0.190 | -0.027 | 0.051 | 0.033 | 0.785 |
| | | | | $\rho = 0.9$ | | | | |
| $\widehat{\rho}^{M}$ | -0.562 | 0.567 | 0.562 | 0.000 | -0.204 | 0.207 | 0.204 | 0.000 |
| $\widehat{\rho}^{M}_{DD}$ | -0.488 | 0.494 | 0.488 | 0.000 | -0.175 | 0.179 | 0.175 | 0.000 |
| $\widehat{\rho}^{M}_{CF}$ | -0.026 | 0.177 | 0.106 | 0.720 | 0.008 | 0.075 | 0.045 | 0.486 |
| $\widehat{\rho}^{E}$ | -0.562 | 0.568 | 0.562 | 0.000 | -0.204 | 0.207 | 0.204 | 0.000 |
| $\widehat{\rho}^{E}_{DD}$ | -0.487 | 0.494 | 0.487 | 0.000 | -0.175 | 0.179 | 0.175 | 0.000 |
| $\widehat{\rho}^{E}_{CF}$ | -0.043 | 0.161 | 0.090 | 0.719 | 0.008 | 0.073 | 0.045 | 0.449 |
| $\widehat{\rho}^{M}_{BCDD}$ | -0.441 | 0.510 | 0.442 | 0.022 | -0.195 | 0.199 | 0.195 | 0.000 |
| $\widehat{\rho}^{M}_{BCCF}$ | -0.227 | 0.249 | 0.227 | 0.029 | -0.063 | 0.074 | 0.063 | 0.198 |

Monte Carlo results for dynamic panel simulation. Results are for the autoregressive parameter in the dynamic panel model where the true value is .6 in the upper panel and .9 in the lower panel. Bias, root mean squared error (RMSE), mean absolute deviation (MAD), and coverage probabilities of 95% level interval estimates (CP(.95)) are reported. $\widehat{\rho}^{M}_{DD}$, $\widehat{\rho}^{M}_{CF}$, $\widehat{\rho}^{E}_{DD}$, and $\widehat{\rho}^{E}_{CF}$ are Bayesian estimators with bias-reducing priors. $\widehat{\rho}^{M}$ is the MLE, and $\widehat{\rho}^{E}$ is the posterior mean under a flat prior. The remaining estimators are bias-corrected maximum likelihood estimators. The number of simulations is 1000.