Econometrica, Vol. 81, No. 5 (September, 2013), 1805–1849

RISK OF BAYESIAN INFERENCE IN MISSPECIFIED MODELS, AND THE SANDWICH COVARIANCE MATRIX

BY ULRICH K. MÜLLER[1]

It is well known that, in misspecified parametric models, the maximum likelihood estimator (MLE) is consistent for the pseudo-true value and has an asymptotically normal sampling distribution with "sandwich" covariance matrix. Also, posteriors are asymptotically centered at the MLE, normal, and of asymptotic variance that is, in general, different than the sandwich matrix. It is shown that due to this discrepancy, Bayesian inference about the pseudo-true parameter value is, in general, of lower asymptotic frequentist risk when the original posterior is substituted by an artificial normal posterior centered at the MLE with sandwich covariance matrix. An algorithm is suggested that allows the implementation of this artificial posterior also in models with high dimensional nuisance parameters which cannot reasonably be estimated by maximizing the likelihood.

KEYWORDS: Posterior variance, quasi-likelihood, pseudo-true parameter value, interval estimation.

1. INTRODUCTION

EMPIRICAL WORK IN ECONOMICS RELIES more and more on Bayesian inference, especially in macroeconomics. For simplicity and computational tractability, applied Bayesian work typically makes strong parametric assumptions about the likelihood. The great majority of Bayesian estimations of dynamic stochastic general equilibrium (DSGE) models and VARs, for instance, assume Gaussian innovations. Such strong parametric assumptions naturally lead to a concern about potential misspecification.

This paper formally studies the impact of model misspecification on the quality of standard Bayesian inference, and suggests a superior mode of inference based on an artificial "sandwich" posterior. To fix ideas, consider the linear regression

$$y_i = z_i'\theta + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where the fitted model treats ε_i as independent and identically distributed (i.i.d.) $\mathcal{N}(0, 1)$ independent of the fixed regressors z_i. The parameter of interest is the population regression coefficient θ, and the (improper) prior den-

[1]I thank Harald Uhlig and three anonymous referees for numerous helpful comments and suggestions. I also benefitted from comments by Jun Yu, Yong Li, Andriy Norets, and Chris Sims, as well as seminar participants at the Greater New York Metropolitan Area Econometrics Colloquium, the North American Winter Meeting of the Econometric Society, the Tsinghua International Conference in Econometrics, Harvard/MIT, LSE, NYU, Oxford, Northwestern, Princeton, Rice, UBC, UCL, and Virginia.

 DOI: 10.3982/ECTA9097

sity p of θ is constant $p(\theta) = 1$. The model ("M") log-likelihood is exactly quadratic around the maximum likelihood estimator (MLE) $\hat{\theta}$,

$$(2) \qquad L_{Mn}(\theta) = C - \frac{1}{2} n(\theta - \hat{\theta})' \Sigma_M^{-1} (\theta - \hat{\theta}),$$

where $\Sigma_M = (n^{-1} \sum_{i=1}^n z_i z_i')^{-1}$ and C is a generic constant. With the flat prior on θ, the posterior density has the same shape as the likelihood, that is, the posterior distribution is $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_M)$. Now suppose the fitted model is misspecified, because the Gaussian innovations ε_i are, in fact, heteroskedastic $\varepsilon_i \sim \mathcal{N}(0, \kappa(z_i))$. The sampling distribution of $\hat{\theta}$ then is

$$(3) \qquad \hat{\theta} \sim \mathcal{N}(\theta, \Sigma_S / n), \quad \Sigma_S = \Sigma_M V \Sigma_M,$$

where $V = n^{-1} \sum_{i=1}^n \kappa(z_i) z_i z_i'$. Note that, under correct specification $\kappa(z_i) = 1$, the "sandwich" covariance matrix Σ_S reduces to Σ_M via $V = \Sigma_M^{-1}$. But, in general, misspecification leads to a discrepancy between the sampling distribution of the MLE $\hat{\theta}$ and the shape of the model likelihood: the log-likelihood (2) is exactly *as if* it was based on the single observation $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma_M / n)$, whereas the actual sampling distribution of $\hat{\theta}$ is as in (3). In other words, the model likelihood does not correctly reflect the sample information about θ contained in $\hat{\theta}$. This suggests that one obtains systematically better inference about θ by replacing the original log-likelihood L_{Mn} by the artificial sandwich log-likelihood

$$(4) \qquad L_{Sn}(\theta) = C - \frac{1}{2} n(\theta - \hat{\theta})' \Sigma_S^{-1} (\theta - \hat{\theta}),$$

which yields the "sandwich posterior" $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_S / n)$. Systematically better here is meant in a classical decision theory sense: the model and sandwich posterior distributions are employed to determine the posterior expected loss minimizing action for some given loss function. Sandwich posterior inference is then defined as superior to original Bayesian inference if it yields lower average realized losses over repeated samples, that is, if it results in decisions of lower frequentist risk.

In general, of course, log-likelihoods are not quadratic, the sampling distribution of the MLE is not Gaussian, and priors are not necessarily flat. But the seminal results of Huber (1967) and White (1982) showed that, in large samples, the sampling distribution of the MLE in misspecified models is centered on the Kullback–Leibler divergence minimizing pseudo-true parameter value and, to first asymptotic order, it is Gaussian with sandwich covariance matrix. The general form of the sandwich matrix involves both the second derivative of the log-likelihood and the variance of the scores, and can be consistently estimated under weak regularity conditions. Similarly, also the asymptotic behavior of the posterior in misspecified parametric models is well understood:

The variation in the likelihood dominates the variation in the prior, leading to a Gaussian posterior centered at the MLE and with covariance matrix equal to the inverse of the second derivative of the log-likelihood. See, for instance, Chapter 11 of Hartigan (1983), Chapter 4.2 and Appendix B in Gelman, Carlin, Stern, and Rubin (2004), or Chapter 3.4 of Geweke (2005) for textbook treatments. The small sample arguments above thus apply at least heuristically to large sample inference about pseudo-true values in general parametric models.

The main point of this paper is to formalize this intuition: Under suitable conditions, the large sample frequentist risk of decisions derived from the sandwich posterior $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_S/n)$ is (weakly) smaller than the risk of decisions derived from the model posterior. This result does not imply that sandwich posterior inference yields the lowest possible asymptotic risk; in fact, only in rather special cases does the sandwich posterior correspond to the posterior computed from the correct model, even asymptotically. But with the correct model unknown and difficult to specify, sandwich posterior inference constitutes a pragmatic improvement for Bayesian inference in parametric models under potential misspecification. We discuss an implementation that is potentially suitable also for models with a high dimensional parameter in the context of a factor model in Section 6.1 below.

It is important to keep in mind that the pseudo-true parameter of the misspecified model must remain the object of interest for sandwich posterior inference to make sense.[2] The pseudo-true parameter is jointly determined by the true data generating process and the fitted model. For instance, with a substantive interest in the population regression coefficient in (1), an assumption of Gaussian ε_i generally leads to a consistent MLE as long as $E[z_i \varepsilon_i] = 0$. In contrast, if the fitted model assumes ε_i to be mean-zero mixtures of normals independent of the regressors, say, then the pseudo-true value does not, in general, equal the population regression coefficient. In this setting, the ostensibly weaker assumption of a mixture of normals distribution for ε_i thus yields *less* robust inference in this sense, and it is potentially attractive to address potential misspecification of the baseline Gaussian model with sandwich posterior inference instead. Section 5.1 below provides numerical evidence on this issue.

One might also question whether losses in decision problems exclusively concern the value of parameters. For instance, the best conditional prediction of y given z in the linear regression (1) is, in general, a function of the conditional distribution of $\varepsilon | z$, and not simply a function of the population regression coefficient θ (unless the loss function is quadratic). At the same time, most "decisions" in Bayesian applied work concern the description of uncertainty about model parameters (and functions of model parameters, such as impulse responses) by two-sided equal-tailed posterior probability intervals. In

[2]Also, Royall and Tsou (2003) and Freedman (2006) stressed that pseudo-true parameters do not necessarily have an interesting interpretation.

a correctly specified model, these intervals are optimal actions relative to a loss function that penalizes long and mis-centered intervals. Under misspecification with $\Sigma_S \neq \Sigma_M$, the sandwich posterior two-sided equal-tailed intervals are a systematically better description of parameter uncertainty in the sense that they are of lower asymptotic risk under this loss function. In the empirical illustration in Section 6.2, we find that the sandwich posterior implies substantially more uncertainty about model parameters in a three-equation DSGE model fitted to postwar U.S. data compared to standard Bayesian inference.

The relatively closest contribution in the literature to the results of this paper seems to be a one-page discussion in Royall and Tsou (2003). They considered Stafford's (1996) robust adjustment to the (profile) likelihood, which raises the original likelihood to a power such that, asymptotically, the inverse of the second derivative of the resulting log-likelihood coincides with the sampling variance of the scalar (profile) MLE to first order. In their Section 8, Royall and Tsou verbally discussed asymptotic properties of posteriors based on the adjusted likelihood, which is equivalent to the sandwich likelihood studied here for a scalar parameter of interest. They accurately noted that the posterior based on the adjusted likelihood is "correct" if the MLE in the misspecified model is asymptotically identical to the MLE of a correctly specified model, but went on to mistakenly claim that otherwise, the posterior based on the adjusted likelihood is conservative in the sense of overstating the variance. See Section 4.4 below for further discussion.

Since the sandwich posterior is a function of the MLE and its (estimated) sampling variance only, the approach of this paper is also related to the literature that constructs robust, "limited information" likelihoods from a statistic, such as a GMM estimator. The Gaussianity of the posterior can then be motivated by the approximately Gaussian sampling distribution of the estimator, as in Pratt, Raiffa, and Schlaifer (1965, Chapter 18.4), Doksum and Lo (1990), and Kwan (1999), or by entropy arguments, as in Zellner (1997) and Kim (2002). Similarly, Boos and Monahan (1986) suggested inversion of the bootstrap distribution of $\hat{\theta}$ to construct a likelihood for θ. The contribution of the current paper relative to this literature is the asymptotic decision theoretic analysis, as well as the comparison to standard Bayesian inference based on the model likelihood.

The remainder of the paper is organized as follows. Section 2 considers in detail the simple setting where the log-likelihood is exactly quadratic and the sampling distribution of the MLE is exactly Gaussian. Section 3 provides the heuristics for the more general large sample result. The formal asymptotic analysis is in Section 4. Sections 5 and 6 contain Monte Carlo results and an empirical application in models with a small and large number of parameters, respectively, and a discussion of implementation issues for sandwich posterior inference. Section 7 concludes. Replication files for the simulations and empirical examples may be found in Müller (2013).

2. ANALYSIS WITH GAUSSIAN MLE AND QUADRATIC LOG-LIKELIHOOD

2.1. *Set-up*

The salient features of the inference regression problem of the Introduction are that the log-likelihood of the fitted model L_{Mn} is exactly quadratic as in (2), and that the sampling distribution of $\hat{\theta}$ in the misspecified model is exactly Gaussian as in (3). In this section, we will assume that Σ_M and Σ_S are constant across samples and do not depend on θ_0. Also, we assume $\Sigma_S \neq \Sigma_M$ to be known—in practice, of course, Σ_S will need to be estimated.

If the misspecification is ignored, then Bayes inference about $\theta \in \mathbb{R}^k$ leads to a posterior that is proportional to the product of the prior p and the likelihood (2). Specifically, given a set of actions $\mathcal{A}$ and a loss function $\ell : \Theta \times \mathcal{A} \mapsto [0, \infty)$, the Bayes action $a \in \mathcal{A}$ minimizes

$$\int \ell(\theta, a)\phi_{\Sigma_M/n}(\theta - \hat{\theta})p(\theta)\,d\theta, \tag{5}$$

where ϕ_Σ is the density of a mean-zero normal with variance Σ. In contrast, taking the sandwich log-likelihood (4) as the sole basis for the data information about θ leads to a posterior expected loss of action a proportional to

$$\int \ell(\theta, a)\phi_{\Sigma_S/n}(\theta - \hat{\theta})p(\theta)\,d\theta. \tag{6}$$

With $\Sigma_M \neq \Sigma_S$, (5) and (6) are different functions of a. Thus, with the same data, basing Bayes inference on the original log-likelihood L_{Mn} will generically lead to a different action than basing Bayes inference on the sandwich log-likelihood L_{Sn}. Denote by $\mathcal{D}$ the set of functions $\Theta \mapsto \mathcal{A}$ that associate a realization of $\hat{\theta}$ with a particular action. Note that, with Σ_M and Σ_S fixed over samples and independent of θ, the actions that minimize (5) and (6) are elements of $\mathcal{D}$. For future reference, denote these functions by d^*_M and d^*_S, respectively. The *frequentist risk* of decision d is

$$r(\theta, d) = E_\theta\big[\ell\big(\theta, d(\hat{\theta})\big)\big] = \int \ell\big(\theta, d(\hat{\theta})\big)\phi_{\Sigma_S/n}(\hat{\theta} - \theta)\,d\hat{\theta}.$$

Note that $r(\theta, d)$ involves the density $\phi_{\Sigma_S/n}$, reflecting that the sandwich covariance matrix Σ_S describes the variability of $\hat{\theta}$ over different samples. The aim is a comparison of $r(\theta, d^*_M)$ and $r(\theta, d^*_S)$.

2.2. *Bayes Risk*

Since frequentist risk is a function of the true value θ, typically decisions have risk functions that cross, so that no unambiguous ranking can be made.

A scalar measure of the desirability of a decision d is *Bayes risk* R, the weighted average of frequentist risk r with weights equal to some probability density η:

$$R(\eta, d) = \int r(\theta, d)\eta(\theta)\, d\theta.$$

Note that with $\eta = p$, we can interchange the order of integration and obtain

$$R(p, d) = \int \int \ell\big(\theta, d(\hat{\theta})\big)\phi_{\Sigma_S/n}(\theta - \hat{\theta})\, p(\theta)\, d\theta\, d\hat{\theta}.$$

If $d_S^* \in \mathcal{D}$ is the decision that minimizes posterior expected loss (6) for each observation $\hat{\theta}$, so that d_S^* satisfies

$$\begin{aligned}&\int \ell\big(\theta, d_S^*(\hat{\theta})\big)\phi_{\Sigma_S/n}(\theta - \hat{\theta})\, p(\theta)\, d\theta\\ &\quad= \inf_{a\in\mathcal{A}} \int \ell(\theta, a)\phi_{\Sigma_S/n}(\theta - \hat{\theta})\, p(\theta)\, d\theta,\end{aligned}$$

then d_S^* also minimizes Bayes risk $R(p, d)$ over $d \in \mathcal{D}$, because minimizing the integrand at all points is sufficient for minimizing the integral. Thus, by construction, d_S^* is the systematically best decision in this weighted average frequentist risk sense.

In contrast, the decision d_M^* that minimizes posterior expected loss (5) computed from the original, misspecified likelihood satisfies

$$\begin{aligned}&\int \ell\big(\theta, d_M^*(\hat{\theta})\big)\phi_{\Sigma_M/n}(\theta - \hat{\theta})\, p(\theta)\, d\theta\\ &\quad= \inf_{a\in\mathcal{A}} \int \ell(\theta, a)\phi_{\Sigma_M/n}(\theta - \hat{\theta})\, p(\theta)\, d\theta.\end{aligned}$$

Clearly, by the optimality of d_S^*, $R(p, d_M^*) \geq R(p, d_S^*)$, and potentially $R(p, d_M^*) > R(p, d_S^*)$.

2.3. *Bayes Risk With a Flat Prior*

Now suppose the prior underlying the posterior calculations (5) and (6) is improper and equal to Lebesgue measure, $p(\theta) = 1$. The shape of the posterior is then identical to the shape of the likelihood, so that, for inference based on the log-likelihood L_{Jn}, $J = M, S$, the posterior becomes

$$\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_J/n) \tag{7}$$

for each realization of $\hat{\theta}$, and the decisions d_J^* satisfy

$$\int \ell\big(\theta, d_J^*(\hat{\theta})\big)\phi_{\Sigma_J/n}(\theta - \hat{\theta})\, d\theta = \inf_{a\in\mathcal{A}} \int \ell(\theta, a)\phi_{\Sigma_J/n}(\theta - \hat{\theta})\, d\theta. \tag{8}$$

Proceeding as in the last subsection, we obtain $R(p, d_M^*) \geq R(p, d_S^*)$, provided $R(p, d_S^*)$ exists.

What is more, if a probability density function $\eta(\theta)$ does not vary much relative to $\ell(\theta, d(\hat{\theta}))\phi_{\Sigma_S/n}(\theta - \hat{\theta})$ for any $\hat{\theta} \in \mathbb{R}^k$, one would expect that

$$R(\eta, d) = \int\int \ell(\theta, d(\hat{\theta}))\phi_{\Sigma_S/n}(\theta - \hat{\theta})\eta(\theta)\, d\theta\, d\hat{\theta} \approx \int\int \ell(\theta, d(\hat{\theta}))\phi_{\Sigma_S/n}(\theta - \hat{\theta})\, d\theta \eta(\hat{\theta})\, d\hat{\theta}. \tag{9}$$

This suggests that also $R(\eta, d_M^*) \geq R(\eta, d_S^*)$, with a strict inequality if the posterior expected loss minimizing action (8) is different for $J = M, S$.

2.4. *Loss Functions*

The results of this paper are formulated for general decision problems and loss functions. To fix ideas, however, it is useful to introduce some specific examples. Many of these examples concern a decision about a scalar element of the $k \times 1$ vector θ, which we denote by $\theta_{(1)}$, and $\hat{\theta}_{(1)}$ is the corresponding element of $\hat{\theta}$. In the following, d_J^* is the posterior expected loss minimizing decision under a flat prior, that is, relative to the posterior $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_J/n)$.

Estimation under quadratic loss: $A = \mathbb{R}$, $\ell(\theta, a) = (\theta_{(1)} - a)^2$, and $d_J^*(\hat{\theta}) = \hat{\theta}_{(1)}$. Under this standard symmetric loss function, the estimator does not depend on the variance. Thus, sandwich posterior inference trivially has the same risk as inference based on the model likelihood.

The results of this paper are more interesting for decision problems where the best rule is a function of the variance Σ_J. This is naturally the case for set estimation problems, but also holds for point estimation problems under asymmetric loss.

Estimation under linex loss: $A = \mathbb{R}$, $\ell(\theta, a) = \exp[b(\theta_{(1)} - a)] - b(\theta_{(1)} - a) - 1$, $b \neq 0$, and $d_J^*(\hat{\theta}) = \hat{\theta}_{(1)} + \frac{1}{2} b \Sigma_{J(1,1)}/n$, where $\Sigma_{J(1,1)}$ is the $(1, 1)$ element of Σ_J. For $b > 0$, linex loss is relatively larger if $\theta_{(1)} - a$ is positive, so the optimal decision tilts the estimator toward larger values, and the optimal degree of this tilting depends on the variability of $\hat{\theta}$.

Interval estimation problem: $A = (a_l, a_u) \in \mathbb{R}^2$, $a_l \leq a_u$, $\ell(\theta, a) = a_u - a_l + c(\mathbf{1}[\theta_{(1)} < a_l](a_l - \theta_{(1)}) + \mathbf{1}[\theta_{(1)} > a_u](\theta_{(1)} - a_u))$, $d_J^*(\hat{\theta}) = [\hat{\theta}_{(1)} - m_J^*, \hat{\theta}_{(1)} + m_J^*]$ with m_J^* the $1 - c^{-1}$ quantile of $\mathcal{N}(0, \Sigma_{J(1,1)}/n)$.[3] This loss function was already mentioned in the Introduction. It rationalizes the reporting of two-sided equal-tailed posterior probability intervals, which is the prevalent method of reporting parameter uncertainty in Bayesian studies. This decision problem is therefore the leading example for the relevance of the results of this paper. The

[3]See Theorem 5.78 of Schervish (1995) for this form of d_J^*.

Monte Carlo and empirical results in Sections 5 and 6 below rely heavily on this loss function.

Multivariate set estimation problem: $A = \{\text{all Borel subsets of } \mathbb{R}^k\}$, $\ell(\theta, a) = \mu_L(a) + c\mathbf{1}[\theta \notin a]$, where $\mu_L(a)$ is the Lebesgue measure of the set $a \subset \mathbb{R}^k$, and $d_J^*(\hat{\theta}) = \{\theta : \phi_{\Sigma_J/n}(\theta - \hat{\theta}) \geq 1/c\}$. This may be viewed as multivariate generalization of the interval estimation problem, and, in general, leads to the reporting of the highest posterior density set. Since the posterior $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_J/n)$ is symmetric and unimodal, it yields optimal decisions of the same form as the interval estimation problem for $k = 1$.

Estimation with an indicator of its precision: $A = \mathbb{R} \times \{0, 1\}$, $a = (a_E, a_P)$, $\ell(\theta, a) = (1 + a_P c_P)\mathbf{1}[|\theta_{(1)} - a_E| > c_D] + (1 - a_P)(1 - c_P)\mathbf{1}[|\theta_{(1)} - a_E| \leq c_D]$, where $c_D > 0$ and $0 < c_P < 1$, and $d_J^*(\hat{\theta}) = (\hat{\theta}_{(1)}, \mathbf{1}[\int_{-c_D}^{c_D} \phi_{\Sigma_J/n}(\theta_{(1)} - \hat{\theta}_{(1)})\, d\theta_{(1)} \geq c_P])$. The problem here is to jointly decide about the value of $\theta_{(1)}$ and whether its guess a_E is within c_D of the true value.

The best decisions in the last four decision problems are functions of Σ_J. This suggests that in these problems, $R(\eta, d_M^*) > R(\eta, d_S^*)$, at least for sufficiently vague η. A more precise statement can be made by exploiting an invariance property.

DEFINITION 1: A loss function $\ell : \Theta \times \mathcal{A} \mapsto \mathbb{R}$ is *invariant* if, for all $\theta \in \Theta = \mathbb{R}^k$ and $a \in \mathcal{A}$,

$$\ell(\theta, a) = \ell\big(0, q(a, -\theta)\big),$$

where $q : \Theta \times \mathcal{A} \mapsto \mathcal{A}$ is a *flow*, that is, $q(a, 0) = a$ and $q(q(a, \theta_1), \theta_2) = q(a, \theta_1 + \theta_2)$ for all $\theta_1, \theta_2 \in \Theta$.

It is not hard to see that the loss functions in all five examples satisfy Definition 1; for the interval estimation problem, for instance, $q(a, \theta) = [a_l + \theta_{(1)}, a_u + \theta_{(1)}]$.

2.5. *Risk Under a Flat Prior and an Invariant Loss Function*

If $a_J^*, J = S, M$ minimizes posterior expected loss with $p(\theta) = 1$ after observing $\hat{\theta} = 0$ under an invariant loss function,

$$\int \ell\big(\theta, a_J^*\big)\phi_{\Sigma_J/n}(\theta)\, d\theta = \inf_{a \in \mathcal{A}} \int \ell(\theta, a)\phi_{\Sigma_J/n}(\theta)\, d\theta,$$

then the invariant rule $d_J^*(\hat{\theta}) = q(a_J^*, \hat{\theta})$ minimizes posterior expected loss under the log-likelihood L_{Jn}, since

$$\int \ell\big(\theta, q(a, \hat{\theta})\big)\phi_{\Sigma_J/n}(\theta - \hat{\theta})\, d\theta = \int \ell(\theta, a)\phi_{\Sigma_J/n}(\theta)\, d\theta. \tag{10}$$

Furthermore, for any invariant rule $d(\hat{\theta}) = q(a, \hat{\theta})$,

$$r(\theta, d) = \int \ell\big(\theta, q(a, \hat{\theta})\big)\phi_{\Sigma_S/n}(\hat{\theta} - \theta)\, d\hat{\theta}$$
$$= \int \ell(\hat{\theta}, a)\phi_{\Sigma_S/n}(\hat{\theta})\, d\hat{\theta}.$$

Thus, for $J = S, M$,

$$r\big(\theta, d_J^*\big) = \int \ell\big(\hat{\theta}, a_J^*\big)\phi_{\Sigma_S/n}(\hat{\theta})\, d\hat{\theta}, \tag{11}$$

that is, the small sample frequentist risk $r(\theta, d_J^*)$ of the invariant rule d_J^* is equal to its posterior expected loss (10) with $p(\theta) = 1$, and by definition, d_S^* minimizes both. This is a special case of the general equivalence between posterior expected loss under invariant priors and frequentist risk of invariant rules; see Chapter 6.6 of Berger (1985) for further discussion and references. We conclude that, for each $\theta \in \Theta$, $r(\theta, d_S^*) \leq r(\theta, d_M^*)$, with equality only if the optimal action a_J^* does not depend on the posterior variance Σ_J/n. Thus, for variance dependent invariant decision problems and a flat prior, the small sample risk of d_S^* is uniformly below the risk of d_M^*, and d_M^* is inadmissible. In particular, this holds for all examples in Section 2.4 except for the estimation problem under quadratic loss.

3. HEURISTIC LARGE SAMPLE ANALYSIS

3.1. *Overview*

The arguments in Sections 2.3 and 2.5 were based on (i) a quadratic model log-likelihood; (ii) Gaussianity of the sampling distribution of the MLE $\hat{\theta}$; (iii) a loss function that depends on the center of the sampling distribution of $\hat{\theta}$; (iv) knowledge of the variance Σ_S of the sampling distribution of $\hat{\theta}$; (v) a flat prior. This section reviews standard distribution theory for maximum likelihood estimators and Bernstein–von Mises arguments for misspecified models, which imply these properties to approximately hold in large samples for a wide range of parametric models.

3.2. *Pseudo-True Parameter Values*

Let x_i, $i = 1, \dots, n$ be an i.i.d. sample with density $f(x)$ with respect to some σ-finite measure μ. Suppose a model with density $g(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}^k$, is fitted, yielding a model log-likelihood equal to $L_{Mn}(\theta) = \sum_{i=1}^n \ln g(x_i, \theta)$. If $f(x) \neq g(x, \theta)$ for all $\theta \in \Theta$, then the fitted model $g(x, \theta)$ is misspecified. Let $\hat{\theta}$ be the MLE, $L_{Mn}(\hat{\theta}) = \sup_{\theta \in \Theta} L_n(\theta)$. Since $n^{-1} L_{Mn}(\theta) \xrightarrow{p} E \ln g(x_i, \theta)$

by a uniform law of large numbers, $\hat{\theta}$ will typically be consistent for the value $\theta_0 = \arg\max_{\theta\in\Theta} E\ln g(x_i,\theta)$, where the expectation here and below is relative to the density f.[4] If f is absolutely continuous with respect to g, then

$$(12)\qquad E\ln g(x_i,\theta) - E\ln f(x_i) = -\int f(x)\ln\frac{f(x)}{g(x,\theta)}\,d\mu(x) = -K(\theta),$$

where $K(\theta)$ is the Kullback–Leibler divergence between $f(x)$ and $g(x,\theta)$, so θ_0 is also the Kullback–Leibler minimizing value $\theta_0 = \arg\min_{\theta\in\Theta} K(\theta)$. For some set of misspecified models, this "pseudo-true" value θ_0 sometimes remains the natural object of interest. As mentioned before, the assumption of Gaussian errors in a linear regression model, for instance, yields $\hat{\theta}$ equal to the ordinary least squares estimator, which is consistent for the population regression coefficient θ_0 as long as the errors are not correlated with the regressors. More generally, then, it is useful to define a true model with density $f(x,\theta)$, where, for each $\theta_0\in\Theta$, $K(\theta) = E\ln\frac{f(x_i,\theta_0)}{g(x_i,\theta)} = \int f(x,\theta_0)\ln\frac{f(x,\theta_0)}{g(x,\theta)}\,d\mu(x)$ is minimized at θ_0; that is, the parameter θ in the true model f is, by definition, the pseudo-true parameter value relative to the fitted model $g(x,\theta)$. Pseudo-true values with natural interpretations arise for fitted models in the exponential family, as in Gourieroux, Monfort, and Trognon (1984), and in generalized linear models (see, for instance, Chapters 2.3.1 and 4.3.1 of Fahrmeir and Tutz (2001)). We follow the frequentist quasi-likelihood literature and assume that the object of interest in a misspecified model is this pseudo-true parameter value, so that in the decision problem, the losses ℓ depend on the action taken, and the value of θ_0. This assumption implicitly restricts the extent of the allowed misspecification.

3.3. *Large Sample Distribution of the Maximum Likelihood Estimator*

Let $s_i(\theta)$ be the score of observation i, $s_i(\theta) = \partial\ln g(x_i,\theta)/\partial\theta$, and $h_i(\theta) = \partial s_i(\theta)/\partial\theta'$. Assuming an interior maximum, we have $\sum_{i=1}^n s_i(\hat{\theta}) = 0$, and by a first-order Taylor expansion,

$$(13)\qquad \begin{aligned} 0 &= n^{-1/2}\sum_{i=1}^n s_i(\theta_0) + \left(n^{-1}\sum_{i=1}^n h_i(\theta_0)\right) n^{1/2}(\hat{\theta}-\theta_0) + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^n s_i(\theta_0) - \Sigma_M^{-1} n^{1/2}(\hat{\theta}-\theta_0) + o_p(1), \end{aligned}$$

[4]As shown by Berk (1966, 1970), though, if the argmax is not unique, then $\hat{\theta}$ might not converge at all.

where $\Sigma_M^{-1} = -E[h_i(\theta_0)] = \partial^2 K(\theta)/\partial\theta\,\partial\theta'|_{\theta=\theta_0}$. Invoking a central limit theorem for the mean-zero i.i.d. random variables $s_i(\theta_0)$, we obtain from (13)

$$(14) \qquad n^{1/2}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_S),$$

where $\Sigma_S = \Sigma_M V \Sigma_M$ and $V = E[s_i(\theta_0)s_i(\theta_0)']$. Note that in a correctly specified model, $\Sigma_S = \Sigma_M$ via the information equality $V = \Sigma_M^{-1}$. Further, Σ_S is typically consistently estimated by $\hat{\Sigma}_S = \hat{\Sigma}_M \hat{V} \hat{\Sigma}_M$, where $\hat{\Sigma}_M^{-1} = -n^{-1}\sum_{i=1}^n h_i(\hat{\theta})$ and $\hat{V} = n^{-1}\sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})'$.

3.4. *Large Sample Properties of the Likelihood and Prior*

From a second-order Taylor expansion of $L_{Mn}(\theta)$ around $\hat{\theta}$, we obtain, for all fixed $u \in \mathbb{R}^k$,

$$(15) \qquad \begin{aligned} &L_{Mn}\big(\hat{\theta} + n^{-1/2}u\big) - L_{Mn}(\hat{\theta}) \\ &\quad = n^{-1/2}u'\sum_{i=1}^n s_i(\hat{\theta}) + n^{-1}u'\sum_{i=1}^n h_i(\hat{\theta})u + o_p(1) \\ &\quad \xrightarrow{p} -\frac{1}{2}u'\Sigma_M^{-1}u, \end{aligned}$$

because $\sum_{i=1}^n s_i(\hat{\theta}) = 0$ and $n^{-1}\sum_{i=1}^n h_i(\hat{\theta}) \xrightarrow{p} E[h_i(\theta_0)] = -\Sigma_M^{-1}$. Thus, in large samples, the log-likelihood L_{Mn} is approximately quadratic in the $n^{-1/2}$-neighborhood of its peak $\hat{\theta}$. By (14), $\hat{\theta} - \theta_0 = O_p(n^{-1/2})$, and by a LLN, $n^{-1}L_{Mn}(\theta) - n^{-1}L_{Mn}(\theta_0) \xrightarrow{p} E\ln g(x_i, \theta) - E\ln g(x_i, \theta_0) < 0$ for all $\theta \neq \theta_0$ by the definition of θ_0. Thus, $L_{Mn}(\theta) - L_{Mn}(\hat{\theta})$ with $\theta \neq \theta_0$ diverges to minus infinity with probability converging to 1. This suggests that, in large samples, the log-likelihood is globally accurately approximated by the quadratic function

$$(16) \qquad L_{Mn}(\theta) \approx C - \frac{1}{2}n(\theta - \hat{\theta})'\Sigma_M^{-1}(\theta - \hat{\theta}).$$

Furthermore, for any prior with Lebesgue density p on Θ that is continuous at θ_0, we obtain, for all fixed $u \in \mathbb{R}^k$,

$$p\big(\theta_0 + n^{-1/2}u\big) \to p(\theta_0).$$

Thus, in the relevant $n^{-1/2}$-neighborhood, the prior is effectively flat, and the variation in the posterior density is entirely dominated by the variation in L_{Mn}.

The large sample shape of the posterior then simply reflects the shape of the likelihood $\exp[L_{Mn}(\theta)]$, so that the posterior distribution obtained from L_{Mn} in (16) becomes close to $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_M/n)$.[5]

4. FORMAL LARGE SAMPLE ANALYSIS

4.1. *Overview*

This section develops a rigorous argument for the large sample superiority of sandwich posterior based inference in misspecified models. The heuristics of the last section are not entirely convincing because the sampling distribution of the MLE is only approximately Gaussian; the posterior from the misspecified model is only approximately Gaussian; and the covariance matrix of the MLE often depends on the true parameter value. The main Theorem 1 below also covers mixed asymptotic normal models, where the covariance matrices Σ_M and Σ_S are random even asymptotically.

4.2. *Setup and Basic Assumptions*

The observations in a sample of size n are vectors $x_i \in \mathbb{R}^r$, $i = 1, \dots, n$, with the whole data denoted by $X_n = (x_1, \dots, x_n)$. The model with log-likelihood function $L_{Mn} : \Theta \times \mathbb{R}^{r\times n} \mapsto \mathbb{R}$ is fitted, where $\Theta \subset \mathbb{R}^k$. In the actual data generating process, X_n is a measurable function $D_n : \Omega \times \Theta \mapsto \mathbb{R}^{r\times n}$, $X_n = D_n(\omega, \theta_0)$, where $\omega \in \Omega$ is an outcome in the probability space $(\Omega, \mathfrak{F}, P)$. Denote by P_{n,θ_0} the induced measure of X_n. The true model is parameterized such that θ_0 is pseudo-true relative to the fitted model, that is, $\int L_{Mn}(\theta_0, X)\, dP_{n,\theta_0}(X) = \sup_\theta \int L_{Mn}(\theta, X)\, dP_{n,\theta_0}(X)$ for all $\theta_0 \in \Theta$. The prior on $\theta \in \Theta$ is described by the Lebesgue probability density p, and the data-dependent posterior computed from the (potentially) misspecified model is denoted by Π_n. Let $\hat{\theta}$ be an estimator of θ (in this and the following sections, $\hat{\theta}$ is no longer necessarily equal to the MLE), and let $d_{\mathrm{TV}}(P_1, P_2)$ be the total variation distance between two measures P_1 and P_2. Denote by $\mathcal{P}^k$ the space of positive definite $k \times k$ matrices. We impose the following high-level condition.

CONDITION 1: Under P_{n,θ_0},

(i) $\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \Sigma_S(\theta_0)^{1/2} Z$ with $Z \sim \mathcal{N}(0, I_k)$ independent of $\Sigma_S(\theta_0)$, $\Sigma_S(\theta_0) \in \mathcal{P}^k$ almost surely, and there exists an estimator $\hat{\Sigma}_S \xrightarrow{p} \Sigma_S(\theta_0)$;

(ii) $d_{\mathrm{TV}}(\Pi_n, \mathcal{N}(\hat{\theta}, \Sigma_M(\theta_0)/n)) \xrightarrow{p} 0$, where $\Sigma_M(\theta_0)$ is independent of Z and $\Sigma_M(\theta_0) \in \mathcal{P}^k$ almost surely.

[5]Note that this convergence of the posterior is stronger than the convergence in distribution (14), as the former is based on a convergence of densities, whereas the latter is a convergence of cumulative distribution functions.

For the case of almost surely constant $\Sigma_M(\theta_0)$ and $\Sigma_S(\theta_0)$, primitive conditions that are sufficient for part (i) of Condition 1, with $\hat{\theta}$ equal to the MLE, may be found in White (1982) for the i.i.d. case, and in Domowitz and White (1982) for the non-i.i.d. case. As discussed in Domowitz and White (1982), however, the existence of a consistent estimator $\hat{\Sigma}_S$ becomes a more stringent assumption in the general dependent case (also see Chow (1984) on this point). Part (ii) of Condition 1 assumes that the posterior Π_n computed from the misspecified model converges in probability to the measure of a normal variable with mean $\hat{\theta}$ and variance $\Sigma_M(\theta_0)/n$ in total variation. Sufficient primitive conditions with $\hat{\theta}$ equal to the MLE were provided by Bunke and Milhaud (1998) and Kleijn and van der Vaart (2012) in models with i.i.d. observations. Shalizi (2009) provided general results on the consistency (but not asymptotic normality) of posteriors under misspecification with dependent data, and the general results of Chen (1985) can, in principle, be used to establish part (ii) also in the non-i.i.d. case.

Condition 1 allows $\Sigma_M(\theta_0)$ and $\Sigma_S(\theta_0)$ to be random, so that the following results also apply to locally asymptotic mixed normal (LAMN) models. See, for instance, Jeganathan (1995) for an overview of LAMN theory. Prominent examples in econometrics are regressions with unit root regressors, and explosive autoregressive models.[6] Section 4.5 below provides a set of sufficient assumptions for Condition 1 that cover time series models with potentially random $\Sigma_M(\theta_0)$ and $\Sigma_S(\theta_0)$.

The decision problem consists of choosing the action a from the topological space of possible actions $\mathcal{A}$. The quality of actions is determined by the sample size dependent, measurable loss function $\ell_n : \mathbb{R}^k \times \mathcal{A} \mapsto \mathbb{R}$. (A more natural definition would be $\ell_n : \Theta \times \mathcal{A} \mapsto \mathbb{R}$, but it eases notation if the domain of ℓ_n is extended to $\mathbb{R}^k \times \mathcal{A}$, with $\ell_n(\theta, a) = 0$ for $\theta \notin \Theta$.)

CONDITION 2: $0 \leq \ell_n(\theta, a) \leq \bar{\ell} < \infty$ for all $a \in \mathcal{A}$, $\theta \in \mathbb{R}^k$, $n \geq 1$.

Condition 2 restricts the loss to be nonnegative and bounded. Bounded loss ensures that small probability events only have a small effect on overall risk, which allows precise statements in combination with the weak convergence and convergence in probability assumptions of Condition 1. In practice, many loss functions are not necessarily bounded, but choosing a sufficiently large bound often leads to similar or identical optimal actions. For instance, for the loss functions introduced in Section 2.4, define a corresponding bounded version as $\min(\ell(\theta, a), \bar{\ell})$. Then, at least for large enough $\bar{\ell}$, the Bayes action in the bounded version is identical to the Bayes action in the original version in the estimation problem under quadratic loss and in the set estimation problem,

[6]The $\sqrt{n}$-convergence rate of Condition 1 may be obtained in such models through a suitable rescaling of the data or the parameters.

and they converge to the Bayes action in the original version in the other three problems as $\bar{\ell} \to \infty$.

The motivation for allowing sample size dependent loss functions is not necessarily that more data lead to a different decision problem; rather, this dependence is also introduced out of a concern for the approximation quality of the large sample results. Because sample information about the parameter θ increases linearly in n, asymptotically nontrivial decision problems are those where differences in θ of the order $O(n^{-1/2})$ lead to substantially different losses. With a fixed loss function, this is impossible, and asymptotic results may be considered misleading. For example, in the scalar estimation problem with bounded quadratic loss $\ell_n(\theta, a) = \min((\theta - a)^2, \bar{\ell})$, risk converges to zero for any consistent estimator. Yet, the risk of $\sqrt{n}$-consistent, asymptotically unbiased estimators with smaller asymptotic variance is relatively smaller for large n, and a corresponding formal result is obtained by choosing $\ell_n(\theta, a) = \min(n(\theta - a)^2, \bar{\ell})$.

In the general setting with data $X_n \in \mathbb{R}^{r\times n}$, decisions d_n are measurable functions from the data to the action space, $d_n : \mathbb{R}^{r\times n} \mapsto \mathcal{A}$. Given the loss function ℓ_n and prior p, frequentist risk and Bayes risk of d_n relative to the probability density η are given by

$$r_n(\theta, d_n) = \int \ell_n\big(\theta, d_n(X)\big)\, dP_{n,\theta}(X),$$

$$R_n(\eta, d_n) = \int r_n(\theta, d_n)\eta(\theta)\, d\theta,$$

respectively.

Bayesian decision theory prescribes to choose, for each observed sample X_n, the action that minimizes posterior expected loss. Assuming that this results in a measurable function, we obtain that the Bayes decision $d_{Mn} : \mathbb{R}^{r\times n} \mapsto \mathcal{A}$ satisfies

$$\int \ell_n\big(\theta, d_{Mn}(X_n)\big)\, d\Pi_n(\theta) = \inf_{a\in\mathcal{A}} \int \ell_n(\theta, a)\, d\Pi_n(\theta) \tag{17}$$

for almost all X_n. We will compare the risk of d_{Mn} with the decision rule that is computed from the sandwich posterior

$$\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_S/n). \tag{18}$$

In particular, suppose d_{Sn} satisfies

$$\int \ell_n\big(\theta, d_{Sn}(X_n)\big)\phi_{\hat{\Sigma}_S/n}(\theta - \hat{\theta})\, d\theta = \inf_{a\in\mathcal{A}} \int \ell_n(\theta, a)\phi_{\hat{\Sigma}_S/n}(\theta - \hat{\theta})\, d\theta \tag{19}$$

for almost all X_n. Note that d_{Sn} depends on X_n only through $\hat{\theta}$ and $\hat{\Sigma}_S$.

4.3. *Auxiliary Assumptions*

The formal argument is easier to develop under a stronger-than-necessary condition and with an initial focus on invariant loss functions.

CONDITION 3: For $J = M, S$:

(i) $\Sigma_{J0} = \Sigma_J(\theta_0)$ is nonrandom;

(ii) $\mathcal{A} = \mathbb{R}^m$, and for all $\theta \in \Theta$, $a \in \mathcal{A}$, and $n \geq 1$, $\ell_n(\theta, a) = \tilde{\ell}(u, \tilde{a})$ with $\tilde{\ell}$ invariant in the sense of Definition 1, $u = \sqrt{n}(\theta - \theta_0)$, and $\tilde{a} = \sqrt{n}q(a, -\theta_0)$;

(iii) $\tilde{a}_J^* = \arg\min_{a\in\mathcal{A}} \int \tilde{\ell}(u, a)\phi_{\Sigma_{J0}}(u)\,du$ is unique, and for any sequence of probability distribution G_n on $\mathbb{R}^k$ satisfying $d_{\mathrm{TV}}(G_n, \mathcal{N}(0, \Sigma_{J0})) \to 0$, $\int \tilde{\ell}(u, \tilde{a}_{Gn}^*)\,dG_n(u) = \inf_{a\in\mathcal{A}} \int \tilde{\ell}(u, a)\,dG_n(u)$ implies $\tilde{a}_{Gn}^* \to \tilde{a}_J^*$;

(iv) $\tilde{\ell}: \mathbb{R}^k \times \mathcal{A} \mapsto [0, \bar{\ell}]$ is continuous at $(u, \tilde{a}_J^*)$ for almost all $u \in \mathbb{R}^k$.

Condition 3(ii) assumes a loss function that explicitly focuses on the $\sqrt{n}$-neighborhood of θ_0. The suitably rescaled loss function and actions are denoted by a tilde, and $u = \sqrt{n}(\theta - \theta_0)$ is the local parameter value. Similarly, define $\hat{u} = \sqrt{n}(\hat{\theta} - \theta_0)$, and $\tilde{\Pi}_n$ as the scaled and centered posterior probability measure such that $\tilde{\Pi}_n(A) = \Pi_n(\{\theta : n^{-1/2}(\theta - \hat{\theta}) \in A\})$ for all Borel subsets $A \subset \mathbb{R}^k$. Thus Condition 1 implies $\hat{u} \Rightarrow \Sigma_{S0}^{1/2} Z$ and $d_{\mathrm{TV}}(\tilde{\Pi}_n, \mathcal{N}(0, \Sigma_{M0})) \xrightarrow{p} 0$. Finally, let the tilde also indicate correspondingly recentered and rescaled decisions, $\tilde{d}_n(X_n) = \sqrt{n}q(d_n(X_n), -\theta_0)$.

For an interval estimation problem, for instance, one could set

$$\begin{aligned}(20)\qquad \ell_n(\theta, a) = \min\bigl(&\sqrt{n}\bigl(a_u - a_l + c\mathbf{1}[\theta_{(1)} < a_l](a_l - \theta_{(1)}) \\ &+ c\mathbf{1}[\theta_{(1)} > a_u](\theta_{(1)} - a_u)\bigr), \bar{\ell}\bigr),\end{aligned}$$

where the scaling by $\sqrt{n}$ prevents all reasonable interval estimators to have zero loss asymptotically. The tilde version $\tilde{\ell}(u, \tilde{a})$ of (20) then recovers the sample size independent, bounded version of the loss function introduced in Section 2.4. Correspondingly, $\tilde{d}_{Mn}(X_n) = (\hat{u}_{(1)} - \kappa_{L\bar{\ell}} m_{Ln}, \hat{u}_{(1)} + \kappa_{R\bar{\ell}} m_{Rn})$, where $\hat{u}_{(1)}$ is the first element of $\hat{u}$, $-m_{Ln}$ and m_{Rn} are the c^{-1} and $(1 - c^{-1})$ quantiles of the first element $u_{(1)}$ of u under $u \sim \tilde{\Pi}_n$, and $\kappa_{L\bar{\ell}}$ and $\kappa_{R\bar{\ell}}$ are correction factors that account for the bound $\bar{\ell}$ and that converge to 1 as as $\bar{\ell} \to \infty$. Similarly, $\tilde{d}_{Sn}(X_n) = (\hat{u}_{(1)} - \kappa_{\bar{\ell}} \hat{m}_n, \hat{u}_{(1)} + \kappa_{\bar{\ell}} \hat{m}_n)$, where $\hat{m}_n$ is the $(1 - c^{-1})$ quantile of $u_{(1)}$ under $u \sim \mathcal{N}(0, \hat{\Sigma}_S)$. It can be shown that Condition 3(iii) and (iv) also hold for loss function (20).

With this notation in place, under Condition 3, the risk of the generic decision d_n under θ_0 is given by

$$(21)\qquad r_n(\theta_0, d_n) = \int \tilde{\ell}\bigl(0, \tilde{d}_n(X)\bigr)\,dP_{n,\theta_0}(X).$$

We want to show that, for $J = S, M$,

$$r_n(\theta_0, d_{Jn}) \to E\big[\tilde{\ell}\big(\Sigma_{S0}^{1/2} Z, \tilde{a}_J^*\big)\big],$$

with the right-hand side identical to the small sample result (11) of Section 2.5.

Now the posterior expected loss of the action $\tilde{d}_{Mn}(X_n)$ satisfies

$$\int \tilde{\ell}\big(u + \hat{u}, \tilde{d}_{Mn}(X_n)\big)\, d\tilde{\Pi}_n(u) = \inf_{a \in \mathcal{A}} \int \tilde{\ell}(u + \hat{u}, a)\, d\tilde{\Pi}_n(u), \tag{22}$$

and by the invariance property of Condition 3(ii), (22) is also equal to

$$\begin{aligned} &\int \tilde{\ell}\big(u, q\big(\tilde{d}_{Mn}(X_n), -\hat{u}\big)\big)\, d\tilde{\Pi}_n(u) \\ &\quad = \inf_{a \in \mathcal{A}} \int \tilde{\ell}\big(u, q(a, -\hat{u})\big)\, d\tilde{\Pi}_n(u) = \inf_{a \in \mathcal{A}} \int \tilde{\ell}(u, a)\, d\tilde{\Pi}_n(u). \end{aligned} \tag{23}$$

Thus, Condition 3(iii) implies that $\tilde{a}_n(X_n) := q(\tilde{d}_{Mn}(X_n), -\hat{u}) \xrightarrow{p} \tilde{a}_M^*$, where the convergence follows from $d_{\mathrm{TV}}(\tilde{\Pi}_n, \mathcal{N}(0, \Sigma_{M0})) \xrightarrow{p} 0$. Therefore, by another application of invariance, Condition 3(iv), and the continuous mapping theorem,

$$\tilde{\ell}\big(0, \tilde{d}_{Mn}(X_n)\big) = \tilde{\ell}\big(-\hat{u}, \tilde{a}_n(X_n)\big) \Rightarrow \tilde{\ell}\big(\Sigma_{S0}^{1/2} Z, \tilde{a}_M^*\big). \tag{24}$$

But convergence in distribution of bounded random variables implies convergence of their expectations, so (21) and (24) imply

$$r_n(\theta_0, d_{Mn}) = \int \tilde{\ell}\big(-\hat{u}, \tilde{a}_n(X)\big)\, dP_{n,\theta_0}(X) \to E\big[\tilde{\ell}\big(\Sigma_{S0}^{1/2} Z, \tilde{a}_M^*\big)\big], \tag{25}$$

as was to be shown. The argument for $r_n(\theta_0, d_{Sn}) \to E[\tilde{\ell}(\Sigma_{S0}^{1/2} Z, \tilde{a}_S^*)]$ is entirely analogous, with the distribution $\mathcal{N}(0, \hat{\Sigma}_S)$ playing the role of $\tilde{\Pi}_n$ and $d_{\mathrm{TV}}(\mathcal{N}(0, \hat{\Sigma}_S), \mathcal{N}(0, \Sigma_{S0})) \xrightarrow{p} 0$ replacing $d_{\mathrm{TV}}(\tilde{\Pi}_n, \mathcal{N}(0, \Sigma_{M0})) \xrightarrow{p} 0$.

While mathematically convenient, Condition 3 is potentially quite restrictive: posterior expected loss minimizing actions are not necessarily unique, even relative to a Gaussian posterior (think of the set estimation problem of Section 2.4), and a generalization to non-Euclidean action spaces $\mathcal{A}$ raises the question of an appropriate metric that underlies the continuity properties in parts (iii) and (iv).

To make further progress, note that as long as $\tilde{a}_M^*$ is expected loss minimizing relative to $\mathcal{N}(0, \Sigma_{M0})$, it satisfies

$$\int \tilde{\ell}\big(u, \tilde{a}_M^*\big) \phi_{\Sigma_{M0}}(u)\, du = \inf_{a \in \mathcal{A}} \int \tilde{\ell}(u, a) \phi_{\Sigma_{M0}}(u)\, du.$$

Thus,

$$(26)\qquad \begin{aligned} 0 &\le \int \tilde{\ell}\big(u, \tilde{a}_n(X_n)\big)\phi_{\Sigma_{M0}}(u)\,du - \int \tilde{\ell}\big(u, \tilde{a}_M^*\big)\phi_{\Sigma_{M0}}(u)\,du \\ &\le \int \tilde{\ell}\big(u, \tilde{a}_n(X_n)\big)\big(\phi_{\Sigma_{M0}}(u)\,du - d\tilde{\Pi}_n(u)\big) \\ &\quad - \int \tilde{\ell}\big(u, \tilde{a}_M^*\big)\big(\phi_{\Sigma_{M0}}(u)\,du - d\tilde{\Pi}_n(u)\big), \end{aligned}$$

where the second inequality follows from (23). But $d_{\text{TV}}(\tilde{\Pi}_n, \mathcal{N}(0, \Sigma_{M0})) \xrightarrow{p} 0$ and Condition 2 imply that, for any sequence $a_n \in \mathcal{A}$,

$$\left|\int \tilde{\ell}(u, a_n)\big(\phi_{\Sigma_{M0}}(u)\,du - d\tilde{\Pi}_n(u)\big)\right| \le \bar{\ell} d_{\text{TV}}\big(\mathcal{N}(0, \Sigma_{M0}), \tilde{\Pi}_n\big) \xrightarrow{p} 0.$$

With the middle expression in (26) bounded below by zero and above by a random variable that converges in probability to zero, we conclude that

$$(27)\qquad \int \tilde{\ell}\big(u, \tilde{a}_n(X_n)\big)\phi_{\Sigma_{M0}}(u)\,du \xrightarrow{p} \int \tilde{\ell}\big(u, \tilde{a}_M^*\big)\phi_{\Sigma_{M0}}(u)\,du.$$

Thus, $\tilde{a}_n(X_n)$ converges in probability to $\tilde{a}_M^*$ in the pseudo-metric $d_{\mathcal{A}}(a_1, a_2) = |\int \tilde{\ell}(u, a_1)\phi_{\Sigma_{M0}}(u)\,du - \int \tilde{\ell}(u, a_2)\phi_{\Sigma_{M0}}(u)\,du|$. For (25) to go through, this convergence must imply the convergence in distribution (24). Thus, it suffices for $\tilde{\ell}$ to be twofold continuous as follows: if a (nonstochastic) sequence of actions a_n comes close to minimizing expected loss relative to $\mathcal{N}(0, \Sigma_{M0})$, then (a) it yields losses close to those of the optimal action $\tilde{a}_M^*$, $\tilde{\ell}(u, a_n) - \tilde{\ell}(u, \tilde{a}_M^*) \to 0$ for almost all $u \in \mathbb{R}^k$, and (b) losses incurred along the sequence $u_n \to u$ are close to those obtained at u, $\tilde{\ell}(u_n, a_n) - \tilde{\ell}(u, a_n) \to 0$ for almost all $u \in \mathbb{R}^k$. Under this assumption, the analysis of d_{Sn} again follows entirely analogously to d_{Mn}.

An additional restrictive feature of Condition 3 is the implicit scalability of actions assumed in part (ii): without a vector space structure on the action space $\mathcal{A}$, $\tilde{a} = \sqrt{n}q(a, -\theta_0)$ is not even defined (think of the estimation-and-signal-of-precision problem of Section 2.4, for instance). In the initial argument leading to (24), Condition 3(ii) was useful to argue for the convergence in probability $\tilde{a}_n(X_n) \xrightarrow{p} \tilde{a}_M^*$. But the refined argument below (27) does not rely on the convergence of actions, but only on the convergence of the implied losses. This makes it possible to do without any scale normalization of actions. A suitable general condition, which also covers random $\Sigma_J(\theta_0)$ as well as loss functions that are not sample size independent functions of $\sqrt{n}(\theta - \theta_0)$, even asymptotically, is as follows.

CONDITION 4: (i) ℓ_n is asymptotically locally invariant at θ_0, that is, there exists a sequence of invariant loss functions ℓ_n^i in the sense of Definition 1 such that

$$\sup_{u\in\mathbb{R}^k}\lim_{n\to\infty}\sup_{a\in\mathcal{A}}\left|\ell_n(\theta_0+u/\sqrt{n},a)-\ell_n^i(\theta_0+u/\sqrt{n},a)\right|=0.$$

For $J=M,S$,

(ii) for sufficiently large n, there exists measurable $a_n^*:\mathcal{P}^k\mapsto\mathcal{A}$ such that, for P-almost all $\Sigma_J(\theta_0)$, $\int\ell_n^i(\theta,a_n^*(\Sigma_J(\theta_0)))\phi_{\Sigma_J(\theta_0)/n}(\theta)\,d\theta=\inf_{a\in\mathcal{A}}\int\ell_n^i(\theta,a)\phi_{\Sigma_J(\theta_0)/n}(\theta)\,d\theta$;

(iii) for P-almost all $\Sigma_J(\theta_0)$ and Lebesgue almost all $u\in\mathbb{R}^k:u_n\to u$ and $\int\ell_n^i(\theta,a_n)\phi_{\Sigma_J(\theta_0)/n}(\theta)\,d\theta-\int\ell_n^i(\theta,a_n^*(\Sigma_J(\theta_0)))\phi_{\Sigma_J(\theta_0)/n}(\theta)\,d\theta\to 0$ for some sequences $a_n\in\mathcal{A}$ and $u_n\in\mathbb{R}^k$ imply $\ell_n^i(u_n/\sqrt{n},a_n)-\ell_n^i(u/\sqrt{n},a_n^*(\Sigma_J(\theta_0)))\to 0$.

It might be instructive to get some sense for Condition 4 by considering two specific loss functions. In the following, assume $\Sigma_J(\theta_0)$ is P-almost surely constant, so that $a_{Jn}^*=a_n^*(\Sigma_J(\theta_0))$, $J=M,S$ is not random.

Consider first the rescaled and bounded loss function (20) of the interval estimation problem. Here $a_{Jn}^*=(-\kappa_{\bar{\ell}}m_J^*/\sqrt{n},\kappa_{\bar{\ell}}m_J^*/\sqrt{n})$, with m_J^* the $1-c$ quantile of the first element of $\mathcal{N}(0,\Sigma_J(\theta_0))$ and $\kappa_{\bar{\ell}}<1$ a correction factor accounting for the bound $\bar{\ell}$ on $\ell_n=\ell_n^i$, and any sequence a_n that satisfies the premise of part (iii) of Condition 4 must satisfy $\sqrt{n}(a_n-a_{Jn}^*)\to 0$. Thus $\ell_n(u_n/\sqrt{n},a_n)-\ell_n(u/\sqrt{n},a_{Jn}^*)\to 0$ for all $u\in\mathbb{R}^k$, so Condition 4 holds.

Second, consider the bounded and scaled set estimation problem of Section 2.4 with $\mathcal{A}=\{$all Borel subsets of $\mathbb{R}^k\}$ and $\ell_n(\theta,a)=\ell_n^i(\theta,a)=\min(n^{k/2}\mu_L(a)+c\mathbf{1}[\theta\notin a],\bar{\ell})$ and $\bar{\ell}$ large. It is quite preposterous, but nevertheless compatible with Condition 1(ii), that the posterior distribution Π_n has a density that essentially looks like $\phi_{\Sigma_M(\theta_0)/n}(\theta-\hat{\theta})$, but with an additional extremely thin (say, of base volume n^{-4}) and very high (say, of height n^2) peak around θ_0, almost surely. If that was the case, then d_{Mn} would, in addition to the highest posterior density region computed from $\phi_{\Sigma_M(\theta_0)/n}(\theta-\hat{\theta})$, include a small additional set of measure n^{-4} that always contains the true value θ_0. The presence of that additional peak induces a substantially different (i.e., lower) risk. It is thus not possible to determine the asymptotic risk of d_{Mn} under Condition 1 in this decision problem, and correspondingly it can be shown that $\ell_n(\theta,a)=\min(n^{k/2}\mu_L(a)+c\mathbf{1}[\theta\notin a],\bar{\ell})$ does not satisfy Condition 4. In the same decision problem with the action space restricted to $\mathcal{A}=\{$all convex subsets of $\mathbb{R}^k\}$, however, the only actions a_n that satisfy the premise of part (iii) in Condition 4 satisfy $d_H(\{u:u/\sqrt{n}\in a_n\},\tilde{a}_J^*)\to 0$, where $\tilde{a}_J^*=\{u:\phi_{\Sigma_J(\theta_0)}(u)\geq 1/c\}$ and d_H is the Hausdorff distance, and $\ell_n^i(u_n/\sqrt{n},a_n)-\ell_n^i(u/\sqrt{n},a_{Jn}^*)\to 0$ holds for all u that are not on the boundary of $\tilde{a}_J^*$.

We now turn to suitable conditions without assuming that ℓ_n is locally invariant. It is then necessary to consider the properties of the random matrices

$\Sigma_M(\theta_0)$ and $\Sigma_S(\theta_0)$ of Condition 1 at more than one point, that is, to view them as stochastic processes $\Sigma_M(\cdot)$ and $\Sigma_S(\cdot)$, indexed by $\theta \in \Theta$.

CONDITION 5: For η an absolutely continuous probability measure on Θ and $J = S, M$,

(i) Condition 1 holds pointwise for η-almost all θ_0 and $\Sigma_J(\cdot)$ is P-almost surely continuous on the support of η;

(ii) for sufficiently large n, there exists a sequence of measurable functions $d_n^* : \Theta \times \mathcal{P}^k \mapsto \mathcal{A}$ so that, for P-almost all $\Sigma_J(\cdot)$, $\int \ell_n(\theta, d_n^*(\theta_0, \Sigma_J(\theta_0)))\phi_{\Sigma_J(\theta_0)/n}(\theta - \theta_0)\,d\theta = \inf_{a\in\mathcal{A}} \int \ell_n(\theta, a)\phi_{\Sigma_J(\theta_0)/n}(\theta - \theta_0)\,d\theta$ for η-almost all $\theta_0 \in \Theta$;

(iii) for η-almost all θ_0, P-almost all $\Sigma_J(\cdot)$, and Lebesgue almost all $u \in \mathbb{R}^k$: $\int \ell_n(\theta, a_n)\phi_{\Sigma_J(\theta_n)/n}(\theta - \theta_n)\,d\theta - \int \ell_n(\theta, d_n^*(\theta_n, \Sigma_J(\theta_n)))\phi_{\Sigma_J(\theta_n)/n}(\theta - \theta_n)\,d\theta \to 0$ and $\sqrt{n}(\theta_n - \theta_0) \to u$ for some sequences $a_n \in \mathcal{A}$ and $\theta_n \in \mathbb{R}^k$ imply $\ell_n(\theta_0, a_n) - \ell_n(\theta_0, d_n^*(\theta_0 + u/\sqrt{n}, \Sigma_J(\theta_0 + u/\sqrt{n}))) \to 0$.

The decisions d_n^* in part (ii) correspond to the optimal decisions in (8) of Section 2.3. Note, however, that in the Gaussian model with a covariance matrix that depends on θ, $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma_J(\theta)/n)$, Bayes actions in (8) would naturally minimize $\int \ell_n(\theta, a)\phi_{\Sigma_J(\theta)/n}(\theta - \hat{\theta})\,d\theta$, whereas the assumption in part (ii) assumes d_n^* to minimize the more straightforward Gaussian location problem with covariance matrix $\Sigma_J(\hat{\theta})/n$ that does not depend on θ. The proof of Theorem 1 below shows that this discrepancy is of no importance asymptotically with the continuity assumption of part (i); correspondingly, the decision d_{Sn} in (19) minimizes Gaussian risk relative to a covariance matrix $\hat{\Sigma}_S$ that does not vary with θ. Part (iii) of Condition 5 is similar to Condition 4(iii) discussed above: If a sequence a_n comes close to minimizing the same risk as $d_n^*(\theta_n, \Sigma_J(\theta_n))$ for some θ_n satisfying $\sqrt{n}(\theta_n - \theta_0) \to u$, then the loss at θ_0 of a_n is similar to the loss of $d_n^*(\theta_0 + u/\sqrt{n}, \Sigma_J(\theta_0 + u/\sqrt{n}))$, at least for Lebesgue almost all u.

4.4. *Main Result and Discussion*

The proof of the following theorem is in the Appendix.

THEOREM 1: (i) *Under Conditions* 1, 2, *and* 4,

$$r_n(\theta_0, d_{Mn}) - E\left[\int \ell_n^i\big(\theta, a_n^*\big(\Sigma_M(\theta_0)\big)\big)\phi_{\Sigma_S(\theta_0)/n}(\theta)\,d\theta\right] \to 0,$$

$$r_n(\theta_0, d_{Sn}) - E\left[\int \ell_n^i\big(\theta, a_n^*\big(\Sigma_S(\theta_0)\big)\big)\phi_{\Sigma_S(\theta_0)/n}(\theta)\,d\theta\right] \to 0.$$

(ii) *Under Conditions* 2 *and* 5,

$$R_n(\eta, d_{Mn}) - E\left[\int\int \ell_n\big(\theta, d_n^*\big(\hat\theta, \Sigma_M(\hat\theta)\big)\big)\phi_{\Sigma_S(\hat\theta)/n}(\theta-\hat\theta)\,d\theta\,\eta(\hat\theta)\,d\hat\theta\right] \to 0,$$

$$R_n(\eta, d_{Sn}) - E\left[\int\int \ell_n\big(\theta, d_n^*\big(\hat\theta, \Sigma_S(\hat\theta)\big)\big)\phi_{\Sigma_S(\hat\theta)/n}(\theta-\hat\theta)\,d\theta\,\eta(\hat\theta)\,d\hat\theta\right] \to 0.$$

1. The results in the two parts of Theorem 1 mirror the discussion of Sections 2.3 and 2.5: For nonrandom Σ_S and Σ_M, the expectation operators are unnecessary, and in large samples, the risk r_n at θ_0 under the (local) invariance assumption, and the Bayes risks R_n of the Bayesian decision d_{Mn} and the sandwich posterior (18) based decision d_{Sn} behave just like in the Gaussian location problem discussed there. In particular, this implies that the decision d_{Sn} is at least as good as d_{Mn} in large samples—formally, the two parts of Theorem 1 yield as a corollary that $\limsup_{n\to\infty}(r_n(\theta_0, d_{Sn}) - r_n(\theta_0, d_{Mn})) \leq 0$ and $\limsup_{n\to\infty}(R_n(\eta, d_{Sn}) - R_n(\eta, d_{Mn})) \leq 0$, respectively. What is more, these inequalities will be strict for many loss functions ℓ_n, since, as discussed in Section 2, decisions obtained with the correct variance often have strictly smaller risk than those obtained from an incorrect assumption about the variance.

2. While asymptotically at least as good as and often better than d_{Mn}, the overall quality of the decision d_{Sn} depends both on the relationship between the misspecified model and the true model, and on how one defines "overall quality." For simplicity, we assume the asymptotic variances to be nonrandom in the following discussion.

First, suppose the data generating process is embedded in a correct parametric model with the same parameter space Θ as the fitted model, and true parameter θ_0. Denote by d_{Cn} and $\hat\theta_C$ the Bayes rule and MLE computed from this correct model (which are, of course, infeasible if the correct model is not known). By the same reasoning as outlined in Section 3, the posterior Π_{Cn} computed from the correct likelihood converges to the distribution $\mathcal{N}(\hat\theta_C, \Sigma_C(\theta_0)/n)$, and $\hat\theta_C$ has the asymptotic sampling distribution $\sqrt{n}(\hat\theta_C - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_C(\theta_0))$. Now if the relationship between the correct model and the misspecified fitted model is such that $\sqrt{n}(\hat\theta_C - \hat\theta) = o_p(1)$, then $\sqrt{n}(\hat\theta - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_S(\theta_0))$ implies $\Sigma_S(\theta_0) = \Sigma_C(\theta_0)$ (even if $\Sigma_M(\theta_0) \neq \Sigma_S(\theta_0)$), and under sufficient smoothness assumptions on ℓ_n, the decisions d_{Sn} and d_{Cn} have the same asymptotic risk. Thus, in this case, d_{Sn} is asymptotically fully efficient. This potential large sample equivalence between a "corrected" posterior and the true posterior if $\sqrt{n}(\hat\theta_C - \hat\theta) = o_p(1)$ was already noted by Royall and Tsou (2003) in the context of Stafford's (1996) adjusted profile likelihood approach.

Second, the sandwich posterior distribution $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_S/n)$ yields the decision with the smallest large sample risk among all artificial posterior distributions centered at $\hat{\theta}$, and d_{Sn} might be considered optimal in this sense. Formally, let Q be a probability measure on $\mathbb{R}^k$, and for given $\bar{\theta} \in \mathbb{R}^k$, let $Q_{\bar{\theta},n}$ be the induced measure of θ when $\sqrt{n}(\theta - \bar{\theta}) \sim Q$. Let d_{Qn} be the decision that satisfies

$$\int \ell_n\big(\theta, d_{Qn}(X_n)\big)\, dQ_{\hat{\theta},n}(\theta) = \inf_{a \in \mathcal{A}} \int \ell_n(\theta, a)\, dQ_{\hat{\theta},n}(\theta).$$

If a^*_{Qn} satisfies $\int \ell^i_n(\theta, a^*_{Qn})\, dQ_{0,n}(\theta) = \inf_{a\in\mathcal{A}} \int \ell^i_n(\theta, a)\, dQ_{0,n}(\theta)$ and Condition 4(iii) also holds for $Q_{0,n}$ and a^*_{Qn} in place of $\mathcal{N}(0, \Sigma_M(\theta_0)/n)$ and $a^*_n(\Sigma_M(\theta_0))$, respectively, then proceeding as in the proof of Theorem 1 yields $r_n(\theta_0, d_{Qn}) - \int \ell^i_n(\theta, a^*_{Qn})\phi_{\Sigma_S(\theta_0)/n}(\theta)\, d\theta \to 0$, so that $\limsup_{n\to\infty}(r_n(\theta_0, d_{Sn}) - r_n(\theta_0, d_{Qn})) \leq 0$. Thus, from a decision theoretic perspective, the best artificial posterior centered at the MLE is the sandwich posterior. This is true whether or not the sandwich posterior is fully efficient by virtue of $\sqrt{n}(\hat{\theta}_C - \hat{\theta}) = o_p(1)$, as discussed above. In contrast, Royall and Tsou (2003) argued on page 402 that "when the adjusted likelihood is not fully efficient, the Bayes posterior distribution calculated by using the adjusted likelihood is conservative in the sense that it overstates the variance (and understates the precision)." This claim seems to stem from the observation that $\Sigma_S(\theta_0) > \Sigma_C(\theta_0)$ when $\sqrt{n}(\hat{\theta}_C - \hat{\theta}) \neq o_p(1)$. But without knowledge of the correct model, $\hat{\theta}_C$ is not feasible, and the *best artificial posterior centered at* $\hat{\theta}$ is the Gaussian sandwich posterior.

Third, some misspecified models yield $\Sigma_S(\theta_0) = \Sigma_M(\theta_0)$, so that no variance adjustment to the original likelihood is necessary. For instance, in the estimation of a linear regression model with Gaussian errors, the MLE for the regression coefficient is the OLS estimator, and the posterior variance $\Sigma_M(\theta_0)$ is asymptotically equivalent to the OLS variance estimator. Thus, as long as the errors are independent of the regressors, the asymptotic variance of the MLE, $\Sigma_S(\theta_0)$, equals $\Sigma_M(\theta_0)$. This is true even though, under non-Gaussian regression errors, knowledge of the correct model would lead to more efficient inference, $\Sigma_C(\theta_0) < \Sigma_S(\theta_0)$. Under the first-order asymptotics considered here, inference based on the original, misspecified model and inference based on sandwich posterior (18) are of the same quality when $\Sigma_S(\theta_0) = \Sigma_M(\theta_0)$.

Finally, d_{Sn} could be an asymptotically optimal decision in some sense because a large sample posterior of the form $\mathcal{N}(\hat{\theta}, \hat{\Sigma}_S/n)$ can be rationalized by some specific prior. In the context of a linear regression model, where the sandwich covariance matrix estimator amounts to White (1980) standard errors, Lancaster (2003) and Szpiro, Rice, and Lumley (2010) provided results in this direction. Also see Schennach (2005) for related results in a General Method of Moments framework.

3. A natural reaction to model misspecification is to enlarge the set of models under consideration, which from a Bayesian perspective simply amounts to

a change of the prior on the model set (although such ex post changes to the prior are not compatible with the textbook decision theoretic justification of Bayesian inference). Model diagnostic checks are typically based on the degree of “surprise” for some realization of a statistic relative to some reference distribution; see Box (1980), Gelman, Meng, and Stern (1996), and Bayarri and Berger (1997) for a review. The analysis here suggests $\hat{\Sigma}_S - \hat{\Sigma}_M$ as a generally relevant statistic to consider in these diagnostic checks, possibly formalized by White's (1982) information matrix equality test statistic.

4. For the problems of parameter interval estimation or set estimation under the losses described in Section 2.4, the practical implication of Theorem 1, part (i) is to report the standard frequentist confidence interval of corresponding level. The large sample equivalence of Bayesian and frequentist description of parameter uncertainty in correctly specified models thus extends to a large sample equivalence of risk minimizing and frequentist description of parameter uncertainty in misspecified models.

4.5. *Justification of Condition 1 With Dependent Data*

For models with dependent observations, such as time series or panel models, it is useful to write the log-likelihood $L_{Mn}(\theta)$ of $X_n = (x_1, \dots, x_n)$ as $L_{Mn}(\theta) = \sum_{t=1}^n l_t(\theta)$, where $l_t(\theta) = L_{Mt}(\theta) - L_{M,t-1}(\theta)$ and $L_{M0}(\theta) = 0$. Define the scores $s_t(\theta) = \partial l_t(\theta)/\partial\theta$ and Hessians $h_t(\theta) = \partial s_t(\theta)/\partial\theta'$. Under regularity conditions about the true model, such as an assumption of $\{x_t\}$ to be stationary and ergodic, a (uniform) law of large numbers can be applied to $n^{-1}\sum_{t=1}^n h_t(\theta)$. Furthermore, note that $\exp[l_t(\theta)]$ is the conditional density of x_t given X_{t-1} in the fitted model. In the correctly specified model, the scores $\{s_t(\theta_0)\}$ thus form a martingale difference sequence (m.d.s.) relative to the information $X_t = (x_1, \dots, x_t)$, $E[s_t(\theta_0)|X_{t-1}] = 0$; cf. Chapter 6.2 of Hall and Heyde (1980). This suggests that, in moderately misspecified models, $\{s_t(\theta_0)\}$ remains an m.d.s., or at least weakly dependent, so that an appropriate central limit theorem can be applied to $n^{-1/2}S_n(\theta_0) = n^{-1/2}\sum_{t=1}^n s_t(\theta_0)$. One would thus expect the heuristic arguments in Section 3 to go through also for time series models. The following theorem provides a corresponding formal result.

THEOREM 2: *If, under P_{n,θ_0},*

(i) *the prior density $p(\theta)$ is continuous and positive at $\theta = \theta_0$;*

(ii) *θ_0 is in the interior of Θ and $\{l_t\}_{t=1}^n$ are twice continuously differentiable in a neighborhood Θ_0 of θ_0;*

(iii) $\sup_{t\leq n} n^{-1/2}\|s_t(\theta_0)\| \xrightarrow{p} 0$, $n^{-1}\sum_{t=1}^n s_t(\theta_0)s_t(\theta_0)' \xrightarrow{p} V(\theta_0)$, *where $V(\theta_0) \in \mathcal{P}^k$ almost surely, and $n^{-1/2}\sum_{t=1}^n s_t(\theta_0) \Rightarrow V(\theta_0)^{1/2}Z$ with $Z \sim \mathcal{N}(0, I_k)$ independent of $V(\theta_0)$;*

(iv) *for all $\epsilon > 0$, there exists $K(\epsilon) > 0$ so that $P_{n,\theta_0}(\sup_{\|\theta-\theta_0\|\geq\epsilon} n^{-1}(L_{Mn}(\theta) - L_{Mn}(\theta_0)) < -K(\epsilon)) \to 1$;*

(v) $n^{-1}\sum_{t=1}^{n}\|h_t(\theta_0)\| = O_p(1)$, *for any null sequence* k_n, $\sup_{\|\theta-\theta_0\|<k_n} n^{-1} \times \sum_{t=1}^{n}\|h_t(\theta) - h_t(\theta_0)\| \xrightarrow{p} 0$ *and* $\sup_{t\leq n,\|\theta-\theta_0\|<k_n} n^{-1}\|h_t(\theta)\| \xrightarrow{p} 0$, *and* $n^{-1} \times \sum_{t=1}^{n} h_t(\theta_0) \xrightarrow{p} -\Sigma_M^{-1}(\theta_0)$, *where* $\Sigma_M(\theta_0) \in \mathcal{P}^k$ *almost surely and* $\Sigma_M(\theta_0)$ *is independent of* Z;

then Condition 1 *holds with* $\hat{\Sigma}_S = \hat{\Sigma}_M\hat{V}\hat{\Sigma}_M$, $\hat{V} = n^{-1}\sum_{t=1}^{n} s_t(\hat{\theta})s_t(\hat{\theta})'$, *and either* (a) $\hat{\theta}$ *equal to the MLE and* $\hat{\Sigma}_M^{-1} = -n^{-1}\sum_{t=1}^{n} h_t(\hat{\theta})$ *or* (b) $\hat{\theta}$ *the posterior median and* $\hat{\Sigma}_M$ *any consistent estimator of the asymptotic variance of the posterior* Π_n.

If also under the misspecified model, $\{s_t(\theta_0)\}$ forms a m.d.s., then the last assumption in part (iii) holds if $\max_{t\leq n} E[s_t(\theta_0)'s_t(\theta_0)] = O(1)$, by Theorem 3.2 of Hall and Heyde (1980) and the so-called Cramer–Wold device. Assumption (iv) is the identification condition employed by Schervish (1995, p. 436) in the context of the Bernstein–von Mises theorem in correctly specified models. It ensures here that evaluation of the fitted log-likelihood at parameter values away from the pseudo-true value yields a lower value with high probability in large enough samples. Assumptions (v) are fairly standard regularity conditions about the Hessians which can be established using the general results in Andrews (1987).

5. APPLICATION: LINEAR REGRESSION

5.1. *Monte Carlo Results*

As a numerical illustration in a low dimensional model, consider a linear regression with coefficient $\theta = (\alpha, \beta)'$ and a single nonconstant regressor w_i,

$$y_i = \alpha + w_i\beta + \varepsilon_i, \quad (y_i, w_i) \sim \text{i.i.d.}, i = 1, \ldots, n. \tag{28}$$

We only consider data generating processes with $E[\varepsilon_i|w_i] = 0$, and assume throughout that the parameter of interest is given by $\beta \in \mathbb{R}$, the population regression slope. If a causal reading of the regression is warranted, interest in β might stem from its usual interpretation as the effect on the *mean* of y_i of increasing w_i by one unit. Also, by construction, $\alpha + w_i\beta$ is the best predictor for $y_i|w_i$ under squared loss. Alternatively, a focus on β might be justified because economic theory implies $E[\varepsilon_i|w_i] = 0$. Clearly, though, one can easily imagine decision problems involving linear models where the natural object of interest is not β; for instance, the best prediction of $y_i|w_i$ under absolute value loss is the median of $y_i|w_i$, which does not coincide with the population regression function $\alpha + w_i\beta$ in general.

We consider six particular data generating processes (DGPs) satisfying (28). In all of them, $w_i \sim \mathcal{N}(0, 1)$. The first DGP is the baseline normal linear model (DMOD) with $\varepsilon_i|w_i \sim \mathcal{N}(0, 1)$. The second model has an error term that is a mixture (DMIX) of two normals where $\varepsilon_i|w_i, s \sim \mathcal{N}(\mu_s, \sigma_s^2)$, $P(s = 1) = 0.8$,

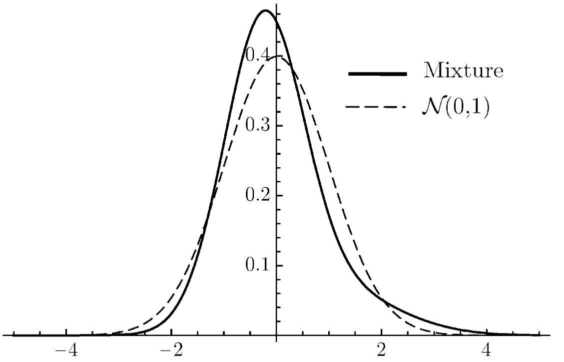

FIGURE 1.—Asymmetric mixture-of-two-normals density.

$P(s=2)=0.2$, $\mu_1=-0.25$, $\sigma_1=0.75$, $\mu_2=1$, and $\sigma_2=\sqrt{1.5}\simeq 1.225$, so that $E[\varepsilon_i^2]=1$. Figure 1 plots the density of this mixture, and the density of a standard normal for comparison. The third model is just like the mixture model, but introduces a conditional asymmetry (DCAS) as a function of the sign of w_i: $\varepsilon_i|w_i,s\sim\mathcal{N}((1-2\cdot\mathbf{1}[w_i<0])\mu_s,\sigma_s^2)$, so that, for $w_i<0$, the distribution of ε_i is the same as the distribution of $-\varepsilon_i$ for $w_i\geq 0$. The final three DGPs are heteroskedastic versions of these homoskedastic DGPs, where $\varepsilon_i|w_i,s=c(0.5+|w_i|)\varepsilon_i^*$, ε_i^* is the disturbance of the homoskedastic DGP, and $c=0.454\ldots$ is the constant that ensures $E[(w_i\varepsilon_i)^2]=1$.

Inference is based on one of the following three methods: First, Bayesian inference with the baseline normal linear regression model (IMOD), where $\varepsilon_i|w_i\sim\mathcal{N}(0,h^{-1})$, with priors $\theta\sim\mathcal{N}(0,100I_2)$ and $3h\sim\chi_3^2$; second, Bayesian inference with a normal mixture linear regression model (IMIX), where $\varepsilon_i|w_i,s\sim\mathcal{N}(\mu_s,(hh_s)^{-1})$, $P(s=j)=\pi_j$, $j=1,2,3$, with priors $\theta\sim\mathcal{N}(0,100I_2)$, $3h\sim\chi_3^2$, $3h_j\sim$ i.i.d. χ_3^2, $(\pi_1,\pi_2,\pi_3)\sim\text{Dirichlet}(3,3,3)$, and $\mu_j|h\sim$ i.i.d. $\mathcal{N}(0,2.5h^{-1})$; third, inference based on the artificial sandwich posterior $\theta\sim\mathcal{N}(\hat\theta,\hat\Sigma_S/n)$ (ISAND), where $\hat\theta=(\hat\alpha,\hat\beta)'$ is the MLE in the baseline normal model (i.e., $\hat\theta$ is the OLS estimator), $\hat\Sigma_S=\hat\Sigma_M\hat V\hat\Sigma_M$, $\hat\Sigma_M=\hat h_n^{-1}(n^{-1}\sum_{i=1}^n z_iz_i')^{-1}$, $\hat V=n^{-1}\hat h_n^2\sum_{i=1}^n z_iz_i'e_i^2$, $\hat h_n^{-1}=n^{-1}\sum_{i=1}^n e_i^2$, $z_i=(1,w_i)'$, and $e_i=y_i-\hat\alpha-w_i'\hat\beta$.

Table I contains the risk of Bayesian inference based on ISAND and IMIX relative to IMOD at $\alpha=\beta=0$ for the scaled and bounded linex loss

$$\ell_n(\theta,a)=\min\bigl(\exp\bigl[2\sqrt{n}(\beta-a)\bigr]-2\sqrt{n}(\beta-a)-1,30\bigr), \tag{29}$$

with $a\in\mathbb{R}$ and scaled and bounded 95% interval estimation loss

$$\ell_n(\theta,a)=\min\bigl(\sqrt{n}\bigl(a_u-a_l+40\cdot\mathbf{1}[\beta<a_l](a_l-\beta)+40\cdot\mathbf{1}[\beta>a_u](\beta-a_u)\bigr),80\bigr), \tag{30}$$

TABLE I

RISK OF DECISIONS ABOUT LINEAR REGRESSION COEFFICIENT[a]

	Homoskedasticity			Heteroskedasticity		
	DMOD	DMIX	DCAS	DMOD	DMIX	DCAS
			Linex Loss, $n = 50$			
ISAND	1.02	1.00	1.09	0.90	0.88	0.97
IMIX	1.02	0.90	1.05	0.91	0.75	1.13
			Linex Loss, $n = 200$			
ISAND	1.01	1.00	1.05	0.87	0.85	0.89
IMIX	1.02	0.85	1.50	0.94	0.72	2.72
			Linex Loss, $n = 800$			
ISAND	1.00	1.00	1.02	0.84	0.85	0.87
IMIX	1.01	0.85	4.02	0.95	0.78	8.14
			Interval Estimation Loss, $n = 50$			
ISAND	1.04	1.02	1.03	0.85	0.84	0.85
IMIX	1.01	0.94	1.03	0.90	0.81	0.96
			Interval Estimation Loss, $n = 200$			
ISAND	1.01	1.01	1.01	0.77	0.75	0.76
IMIX	1.02	0.91	1.25	0.90	0.78	1.86
			Interval Estimation Loss, $n = 800$			
ISAND	1.00	1.00	1.00	0.74	0.74	0.73
IMIX	1.01	0.90	2.66	0.92	0.82	5.64

[a] Data generating processes are in columns, modes of inference in rows. Entries are the risk under linex loss (29) and interval estimation loss (30) relative to risk of standard normal linear regression Bayesian inference (IMOD). Risks are estimated from 10,000 draws for each DGP. The Monte Carlo standard errors for the log of the table entries are between 0.002 and 0.022.

with $a = (a_l, a_u) \in \mathbb{R}^2$, $a_u \geq a_l$, respectively. The bounds are approximately 20 times larger than the median loss for inference using ISAND; unreported simulations show that the following results are quite insensitive to this choice.

In general, IMOD is slightly better than ISAND under homoskedasticity, with a somewhat more pronounced difference in the other direction under heteroskedasticity. This is not surprising, as IMOD is large sample equivalent to inference based on the artificial posterior $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_M/n)$, and $\hat{\Sigma}_M$ is presumably a slightly better estimator of $\Sigma_S(\theta_0)$ than $\hat{\Sigma}_S$ under homoskedasticity, but inconsistent under heteroskedasticity. IMIX performs substantially better than IMOD in the correctly specified homoskedastic mixture model DMIX, but it does very much worse under conditional asymmetry (DCAS) when n is large. It is well known that the OLS estimator achieves the semiparametric efficiency bound in the homoskedastic regression model with $E[\varepsilon_i|w_i] = 0$ (see, for instance, Example 25.28 in van der Vaart (1998) for a textbook exposition), so the lower risk under DMIX has to come at the cost of worse inference in some other DGP. In fact, the pseudo-true value β_0 in the mixture model underlying IMIX under DCAS is *not* the population regression coefficient $\beta = 0$, but a numerical calculation based on (12) shows β_0 to be approximately equal to -0.06. In large enough samples, the posterior for β in this model under DCAS thus concentrates on a nonzero value, and the relative superiority of ISAND is only limited by the bound in the loss functions. Intuitively, under DCAS, IMIX downweighs observations with disturbances that are large in absolute value. Since ε_i is right-skewed for $w_i \geq 0$ and left-skewed for $w_i < 0$, this downweighing tends to occur mostly with positive disturbances when $w_i \geq 0$, and negative disturbances if $w_i < 0$, which leads to a negative bias in the estimation of β.

The much larger risk of IMIX relative to ISAND and IMOD under DCAS suggests that one must be quite sure of the statistical independence of ε_i and w_i before it becomes worthwhile to try to gain efficiency in the non-Gaussian model DMIX. In contrast, the textbook advice seems to favor models with more flexible disturbances as soon as there is substantial evidence of non-Gaussianity. Alternatively, one might, of course, model a potential conditional asymmetry of $\varepsilon_i|w_i$, possibly along the lines recently suggested by Pelenis (2010).

In summary, if the object of interest is the population regression coefficient, then an important property of the normal linear regression model is that the MLE remains consistent whenever the disturbances are mean independent of the regressors. Further, in accordance with Theorem 1, replacing the posterior of this model by the sandwich posterior $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_S/n)$ yields systematically lower risk in misspecified models, at least in medium and large samples.

5.2. *Empirical Illustration*

In Table 14.1 of their textbook, Gelman et al. (2004) reported empirical results on the effect of candidate incumbency on vote shares in congressional elections, using data from 312 contested House of Representatives districts in 1988. The dependent variable is the vote share of the incumbent party, that is, the party that won in 1986. The explanatory variable of interest is an indicator whether the incumbent office holder runs for reelection. The incumbent party (Democratic or Republican) and the vote share of the incumbent party in 1986 are included as controls. Gelman et al. (2004) considered a normal linear regression model with a flat prior on the regression coefficient and the log error variance, so that the posterior mean is exactly equal to the OLS coefficient.

Table II reports posterior mean and standard deviations for IMOD, ISAND, and IMIX in this linear regression, with priors as described in the last subsection, except for $h/100 \sim \chi^2_3$ and a four-dimensional $\mathcal{N}(0, 100I_4)$ prior on the regression coefficients. The IMOD posterior is numerically very close to what was reported in Gelman et al. (2004). The sandwich posterior of the incumbency coefficient has almost the same mean, but the variance is about twice as large. This immediately implies that ISAND results in a substantially different action compared to IMOD in decision problems that seek to describe the uncertainty about the magnitude of the incumbency effect to other political

TABLE II

POSTERIOR MEANS AND STANDARD DEVIATIONS IN INCUMBENCY ADVANTAGE REGRESSION

	IMOD	ISAND	IMIX
Incumbency	0.114	0.114	0.119
	(0.015)	(0.020)	(0.019)
Vote proportion in 1986	0.654	0.654	0.662
	(0.039)	(0.048)	(0.043)
Incumbent party	−0.007	−0.007	−0.007
	(0.004)	(0.004)	(0.004)
Constant	0.127	0.127	0.115
	(0.031)	(0.039)	(0.059)

scientists using the interval or set estimation loss functions of Section 2.4. It is also easy to imagine other decision problems where the difference in uncertainty leads to a different optimal action. For instance, suppose an incumbent candidate credibly threatens the party leader not to run again unless she is made chair of some committee. If the party leader views granting the committee chair as well as losing the district as costly, and the incumbency coefficient is viewed as causal, then there will be a range of beliefs of the party leader about his party's reelection prospects in the district that leads to a different optimal action under IMOD and ISAND.

The posterior mean of the incumbency variable under IMIX is noticeably larger than under IMOD and ISAND. Figure 2 displays kernel estimates of the error density for the two subgroups defined by the incumbency variable. Not only are these two densities of different scale, underlying the difference between the posterior standard deviation of IMOD and ISAND, but also their shapes are quite different. This empirical example thus exhibits the same qualitative properties as the DCAS data generating process of the last subsection.

6. APPLICATION: MODELS WITH A HIGH DIMENSIONAL PARAMETER

6.1. *Monte Carlo Results in a Factor Model*

Consider the following model of the 10 observed time series $\{y_{j,t}\}_{t=1}^n$, $j = 1, \dots, 10$:

$$y_{j,t} = \alpha_j + \beta_j f_t + u_{j,t}, \quad t = 1, \dots, n, \tag{31}$$

where f_t is a scalar unobserved stochastic factor of unit variance $V[f_t] = 1$, β_j is the factor loading of series j, and the idiosyncratic shocks $u_{j,t}$ are mutually independent, independent of $\{f_t\}_{t=1}^n$ and $V[u_{j,t}] = \sigma_j^2$. Suppose the model is estimated under the assumption that $f_t \sim$ i.i.d. $\mathcal{N}(0, 1)$ and $u_{j,t} \sim$ i.i.d. $\mathcal{N}(0, \sigma_j^2)$,

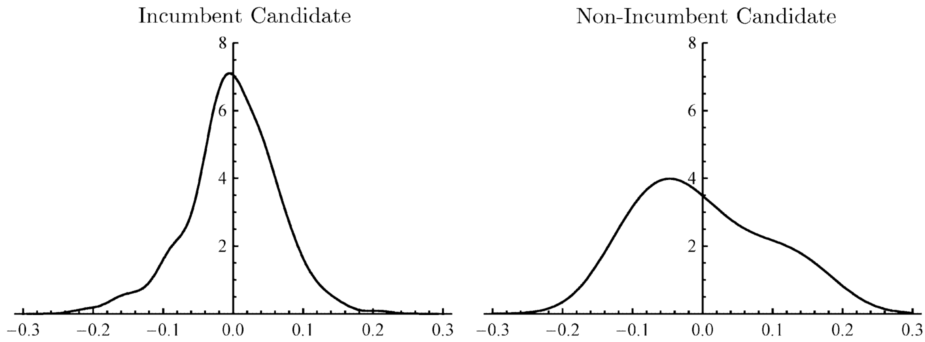

FIGURE 2.—Error densities in incumbency regression conditional on the two values of regressor of interest. Densities are estimated with a Gaussian kernel using Silverman's (1986, p. 48) rule-of-thumb bandwidth.

so that there are three parameters per series and θ is of dimension 30. Call this data generating process DMOD.

We consider two additional data generating processes for (31), under which the fitted model is misspecified. In the first one, DSTDT, $\sqrt{4/3}f_t \sim$ i.i.d. $\mathcal{T}_8$ and $\sqrt{4/3}u_{j,t}/\sigma_j \sim$ i.i.d. $\mathcal{T}_8$, where $\mathcal{T}_8$ is a Student-t distribution with 8 degrees of freedom, and the scaling by $\sqrt{4/3}$ ensures the variances are as in DMOD. In the second one, DAR(1), f_t and $u_{j,t}$ are independent mean-zero stationary Gaussian AR(1) processes with autoregressive coefficient 0.3 and the same variance as under DMOD, so that the estimated model is dynamically misspecified. Under all data generating processes, $\alpha_j = \beta_j = \sigma_j^2 = 1$ for $j = 1, \dots, 10$.

The baseline mode of inference is standard Bayesian inference with a $\mathcal{N}((1,1)', I_2)$ prior on (α_j, β_j) and an independent prior $3\sigma_j^{-2} \sim \chi_3^2$, independent across j. This is implemented via a standard Gibbs sampler.

Given the large number of parameters and the presence of the latent factor f_t, it would be numerically difficult and presumably quite unreliable to implement sandwich posterior inference in a maximum likelihood framework for this (and similar) models. Instead, recall from the discussion in Section 3.4 and Theorem 2 that the posterior distribution obtained from the misspecified likelihood and prior p is approximately $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_M/n)$. The center and scale of the usual Monte Carlo posterior draws can thus be used to construct an appropriate pair $(\hat{\theta}, \hat{\Sigma}_M/n)$. Since the posterior approximation $\mathcal{N}(\hat{\theta}, \Sigma_M/n)$ might not be accurate in the tails, it makes sense to rely on the median, interquartile range, and rank correlations of the posterior as follows: With $Q^{\Pi}(q)$ the element-wise qth quantile of the $k \times 1$ posterior draws, set $\hat{\theta} = Q^{\Pi}(0.5)$ and $\hat{\Sigma}_M/n = \hat{D}\hat{R}\hat{D}$, where $\hat{D} = \text{diag}(Q^{\Pi}(0.75) - Q^{\Pi}(0.25))/1.349$, the i,lth element of the $k \times k$ matrix $\hat{R}$ is equal to $2\sin(\pi\rho_{i,l}/6)$, and $\rho_{i,l}$ is Spearman's rank correlation between the ith and lth element of the posterior draws of θ.[7] The only remaining missing piece for the sandwich posterior $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_S)$ with $\hat{\Sigma}_S = \hat{\Sigma}_M \hat{V} \hat{\Sigma}_M$ now is an estimator $\hat{V}$ of the variance of the scores V.

A natural starting point for the estimation of V is the usual average of outer products of scores, $n^{-1} \sum_{t=1}^{n} s_t(\hat{\theta}) s_t(\hat{\theta})'$, evaluated at the posterior median $\hat{\theta}$. As a small sample correction, it makes sense to add the curvature $-H_p(\theta) = -\partial^2 \ln p(\theta)/\partial\theta\,\partial\theta'$ of the prior density at $\hat{\theta}$,

$$\hat{V} = n^{-1} \sum_{t=1}^{n} s_t(\hat{\theta}) s_t(\hat{\theta})' - n^{-1} H_p(\hat{\theta}). \tag{32}$$

The idea is that the curvature of the posterior as measured by $\hat{\Sigma}_M^{-1}$ is the sum of the curvature of the likelihood, $-n^{-1} \sum_{t=1}^{n} h_t(\theta)$, and the curvature of the prior $-n^{-1} H_p(\theta)$. In a correctly specified model, one would like to have $\hat{\Sigma}_S = \hat{\Sigma}_M \hat{V} \hat{\Sigma}_M$ close to $\hat{\Sigma}_M$—after all, under correct specification, inference based on the exact posterior distribution is small sample Bayes risk minimizing. But without the correction for $\hat{V}$, $\hat{\Sigma}_S$ is systematically smaller than $\hat{\Sigma}_M$, as the variance of the scores only captures the $-n^{-1} \sum_{t=1}^{n} h_t(\theta)$ component of $\hat{\Sigma}_M^{-1}$.[8]

Finally, to avoid the issue of sandwich posterior mass outside Θ, it makes sense to pick a parameterization for θ in which $\Theta = \mathbb{R}^k$. We therefore parameterize the 30 parameters of the factor model (31) as $\{\alpha_j, \beta_j, \ln \sigma_j^2\}_{j=1}^{10}$. Inference based on this form of sandwich posterior is denoted IS-IID. Under potential dynamic misspecification, such as the data generating process DAR(1), a Newey and West (1987)–type HAC variance estimator should replace the simple outerproduct of scores in (32). For the factor model application, we choose a lag-length of 4 and denote the resulting sandwich posterior inference by IS-NW.

Table III reports the risks of the interval estimation problem (30) about α_1, β_1, and σ_1 (note that σ_1 is a nonlinear function of θ). Despite the rather moderate sample sizes, sandwich posterior inference does not lead to large increases in risk under correct specification (i.e., DMOD). At the same time, Student-t innovations favor sandwich inference about the covariance parameters β_1 and σ_1, and the Newey–West (1987)-version of the sandwich matrix estimator also leads to improved risk for the mean parameter α_1 under autocorrelation.

One might wonder why sandwich posterior inference does reasonably well here; after all, when k is large, then n needs to be very large for $\hat{V}$ in (32)

[7] Moran (1948) derived the underlying relationship between the correlation and Spearman's rank correlation of a bivariate normal vector.

[8] Another way of thinking about the correction is as follows: Bayes inference in the Gaussian shift model $Y \sim \mathcal{N}(\theta, \Sigma)$ with prior $\theta \sim \mathcal{N}(0, H_p^{-1})$ is equivalent to Bayes inference with a flat prior and observations (Y, Y_p), where $Y_p \sim \mathcal{N}(\theta, H_p^{-1})$ is independent of Y conditional on θ. The correction term then captures the variance of the score of the additional Y_p observation.

TABLE III

RISK OF INTERVAL ESTIMATION DECISIONS IN FACTOR MODEL[a]

	α_1			β_1			σ_1		
	DMOD	DSTDT	DAR(1)	DMOD	DSTDT	DAR(1)	DMOD	DSTDT	DAR(1)
				$n = 50$					
IS-IDD	1.00	1.00	1.05	1.01	0.96	1.01	1.01	0.98	1.03
IS-NW	1.03	1.02	0.99	1.02	0.99	1.03	1.03	0.99	1.05
				$n = 100$					
IS-IDD	1.00	1.00	1.03	1.00	0.97	1.01	1.00	0.94	1.02
IS-NW	1.02	1.02	0.93	1.01	0.98	1.01	1.02	0.95	1.03
				$n = 200$					
IS-IDD	1.00	1.00	1.02	1.01	0.97	1.01	1.00	0.94	1.00
IS-NW	1.01	1.01	0.90	1.01	0.98	1.00	1.01	0.95	1.00

[a]Data generating processes are in columns, modes of inference in rows. Entries are the risk under interval estimation loss (30) relative to risk of standard Bayesian inference assuming Gaussian innovations in model (31), IMOD. Implementation is as in Table I. Monte Carlo standard errors are between 0.002 and 0.01.

to be an accurate estimator of the $k \times k$ matrix V. But for decision problems that effectively depend on a low dimensional function of θ, such as those in Table III, it is not necessary to estimate the whole matrix V accurately. The (estimated) sandwich posterior for a scalar parameter of interest $\iota'\theta$, say, is given by $\iota'\theta \sim \mathcal{N}(\iota'\hat{\theta}, \iota'\hat{\Sigma}_M \hat{V} \hat{\Sigma}_M \iota)$. Thus, the law of large numbers in (32) only needs to provide a good approximation for the scalar random variables $\iota'\hat{\Sigma}_M s_t(\hat{\theta})$, and moderately large n might well be sufficient for that purpose.

In the simple static factor model (31), it is fairly straightforward to derive the scores $s_t(\hat{\theta})$ from the n observations $x_t = \{y_{j,t}\}_{j=1}^{10}$ analytically. In more complicated models, however, it might not be feasible to integrate out the latent factors in closed form. The following identity might then be useful: Let ξ be the unobserved stochastic component with probability density $p^c(\xi|\theta)$ given θ, and denote by $L_t^c(\theta, \xi)$ the (misspecified) model log-likelihood of $X_t = (x_1, \dots, x_t)$ conditional on ξ (and $L_0^c(\theta, \xi) = 0$). In the factor model, for instance, $\xi = \{f_t\}_{t=1}^n$, $p^c(\xi|\theta)$ does not depend on θ and $L_t^c(\theta, \xi) = -\frac{1}{2}\sum_{s=1}^t \sum_{j=1}^{10} (\ln \sigma_j^2 + (y_{j,s} - \alpha_j - \beta_j f_s)^2/\sigma_j^2)$ up to a constant. The overall model log-likelihood $L_{Mt}(\theta)$ equals $\ln \int \exp[L_t^c(\theta, \xi)] p^c(\xi|\theta)\, d\xi$, so that a calculation yields

$$s_t(\theta) = \int T_t^c(\theta, \xi)\, d\Pi_t^c(\xi|\theta) - \int T_{t-1}^c(\theta, \xi)\, d\Pi_{t-1}^c(\xi|\theta),$$

where $T_t^c(\theta, \xi) = \partial L_t^c(\theta, \xi)/\partial\theta + \partial \ln p^c(\xi|\theta)/\partial\theta$ and $\Pi_t^c(\xi|\theta)$ is the conditional distribution of ξ given (θ, X_t) implied by the fitted model. One can therefore employ n Monte Carlo samplers for ξ using data $X_1, \dots, X_n$ conditional on $\theta = \hat{\theta}$ and compute the posterior averages of the derivatives $T_t^c(\hat{\theta}, \xi)$

to obtain $\{s_t(\hat{\theta})\}_{t=1}^n$. In the factor model example, for the component of $s_t(\hat{\theta})$ corresponding to the derivative with respect to β_i, say, one could take many draws $\{f_s^{(l)}\}_{s=1}^n$, $l = 1, 2, \dots$ of the factor $\{f_s\}_{s=1}^n$ from its conditional distribution given X_t and $\theta = \hat{\theta}$, and compute the simple average of the draws $\partial L_t^c(\theta, \xi)/\partial\beta_i|_{\theta=\hat{\theta}, \xi=\{f_s^{(l)}\}_{s=1}^n} = \sum_{s=1}^t (y_{i,s} - \hat{\alpha}_i - \hat{\beta}_i f_s^{(l)}) f_s^{(l)}/\hat{\sigma}_i^2$. The difference of these averages for t and $t-1$ yields the element in $s_t(\hat{\theta})$ corresponding to β_i.

6.2. *Empirical Illustration*

An important area of applied Bayesian work in economics is the estimation of dynamic stochastic general equilibrium models. For instance, Lubik and Schorfheide (2004; henceforth, LS) estimated a model which, after log-linearization, is given by the three equations

$$y_t = E_t[y_{t+1}] - \tau\big(R_t - E_t[\pi_{t+1}]\big) + g_t, \tag{33}$$

$$\pi_t = \beta E_t[\pi_{t+1}] + \kappa(y_t - z_t), \tag{34}$$

$$R_t = \rho_R R_{t-1} + (1 - \rho_R)\big(\psi_1 \pi_t + \psi_2(y_t - z_t)\big) + \varepsilon_{R,t}, \tag{35}$$

where y_t, π_t, and R_t are percentage deviations from steady state output, inflation, and interest rate, respectively. The steady state of yearly inflation and real interest rates are π^* and r^*, and the quarterly discount factor β is approximated by $\beta = (1 + r^*/100)^{-1/4}$. In addition to the i.i.d. monetary policy shock $\varepsilon_{R,t} \sim (0, \sigma_R^2)$, the two additional shock processes are the demand shock g_t and the productivity shock z_t

$$g_t = \rho_g g_{t-1} + \varepsilon_{g,t}, \quad z_t = \rho_z z_{t-1} + \varepsilon_{z,t}, \tag{36}$$

where $(\varepsilon_{g,t}, \varepsilon_{z,t})$ are i.i.d. with $V[\varepsilon_{g,t}] = \sigma_g^2$, $V[\varepsilon_{z,t}] = \sigma_z^2$, and $E[\varepsilon_{g,t}\varepsilon_{z,t}] = \rho_{zg}\sigma_g\sigma_z$. Let θ be a parameterization of the 13 unknowns of this model.

LS estimated this model under the assumption that the i.i.d. shock process $\varepsilon_t = (\varepsilon_{R,t}, \varepsilon_{g,t}, \varepsilon_{z,t})'$ is Gaussian. This is convenient, since the Kalman filter can then be applied to evaluate the likelihood after casting the linear system (33), (34), and (35) in state space form $Y_t = (y_t, \pi_t, R_t)' = \mu_Y + A(\theta)\xi_t$, $\xi_t = G(\theta)\xi_{t-1} + Q(\theta)\varepsilon_t$. Gaussianity of ε_t, however, is not a defining feature of the model: All white noise processes for ε_t lead to identical second-order properties Y_t, and thus to the same pseudo-true value for θ relative to the Gaussian model. At the same time, the informativeness of the data about θ depends on fourth-order properties of Y_t, which are taken into account by the sandwich posterior, but not by the Gaussian likelihood.

As long as the state space representation is invertible, the state ξ_{t-1} can effectively be computed from $\{Y_s\}_{s=1}^{t-1}$ (at least for t large enough so that the impact of unobserved initial values has died out). Thus, conditional on the true value of θ, the error in the Kalman prediction of Y_t given $\{Y_s\}_{s=1}^{t-1}$ is a linear

TABLE IV

PRIOR AND POSTERIOR IN LUBIK AND SCHORFHEIDE'S (2004) DSGE MODEL[a]

	Prior				95% Posterior Probability Interval		
	trans	shape	mean	stdev	IMOD	IS-IID	IS-NW
ψ_1	$\ln(\psi_1)$	$\mathcal{G}$	1.50	0.25	[1.06, 1.57]	[0.99, 1.72]	[1.00, 1.70]
ψ_2	$\ln(\psi_2)$	$\mathcal{G}$	0.25	0.15	[0.02, 0.32]	[0.03, 0.35]	[0.03, 0.38]
ρ_R	$\ln(\frac{\rho_R}{1-\rho_R})$	$\mathcal{B}$	0.50	0.20	[0.66, 0.79]	[0.64, 0.81]	[0.64, 0.81]
π^*	$\ln(\pi^*)$	$\mathcal{G}$	4.00	2.00	[3.21, 5.23]	[3.30, 5.41]	[3.28, 5.44]
r^*	$\ln(r^*)$	$\mathcal{G}$	2.00	1.00	[1.37, 2.78]	[1.47, 2.90]	[1.40, 3.04]
κ	$\ln(\kappa)$	$\mathcal{G}$	0.50	0.20	[0.16, 0.88]	[0.12, 1.75]	[0.14, 1.56]
τ^{-1}	$\ln(\tau^{-1})$	$\mathcal{G}$	2.00	0.50	[2.03, 4.54]	[1.90, 5.27]	[1.91, 5.22]
ρ_g	$\ln(\rho_g)$	$\mathcal{B}$	0.70	0.10	[0.81, 0.91]	[0.76, 0.93]	[0.77, 0.93]
ρ_z	$\ln(\rho_z)$	$\mathcal{B}$	0.70	0.10	[0.74, 0.86]	[0.70, 0.86]	[0.71, 0.86]
ρ_{gz}	$\ln(\frac{1+\rho_{gz}}{1-\rho_{gz}})$	$\mathcal{N}_{[-1,1]}$	0.00	0.40	[0.38, 0.92]	[0.26, 0.97]	[0.27, 0.97]
ω_R	$\ln(\omega_R)$	$\mathcal{IG}$	0.31	0.16	[0.25, 0.33]	[0.22, 0.36]	[0.21, 0.38]
ω_g	$\ln(\omega_g)$	$\mathcal{IG}$	0.38	0.20	[0.12, 0.21]	[0.11, 0.22]	[0.10, 0.23]
ω_z	$\ln(\omega_z)$	$\mathcal{IG}$	1.00	0.52	[0.83, 1.16]	[0.78, 1.19]	[0.78, 1.20]

[a] $\mathcal{B}$, $\mathcal{G}$, and $\mathcal{N}_{[-1,1]}$, are Beta, Gamma, and Normal (restricted to the $[-1, 1]$ interval) prior distributions, and $\mathcal{IG}$ are Gamma prior distributions on $1/\omega^2$ that imply the indicated mean and standard deviations for ω. The "trans" column specifies the reparameterization underlying the sandwich posterior approximation in $\mathbb{R}^{13}$.

combination of ε_t. With ε_t an m.d.s., this implies that the scores computed from the Kalman filter remain an m.d.s., justifying the estimator $\hat{\Sigma}_S$ described in the previous subsection via Theorem 2.[9]

Following LS, we re-estimate model (33), (34), and (35) on quarterly U.S. data from 1960:I to 1997:IV, so that $n = 132$.[10] Table IV reports the prior and 95% equal-tailed posterior probability intervals from the model implied posterior, and the two sandwich posteriors of the last subsection. Except for the mean parameters π^* and r^*, the sandwich posterior indicates more uncertainty about the model parameters, and often by a substantial amount. A particularly drastic case is the slope of the Phillips curve κ; this is in line with a general fragility of inference about κ across models and specifications discussed by Schorfheide (2008).

The differences between model and sandwich posterior probability intervals are driven by severe departures from Gaussianity: The Kalman forecast errors for y_t, π_t, and R_t display an excess kurtosis of 1.28, 0.82, and 7.61, respectively.[11]

[9]The scores $\{s_t(\hat{\theta})\}$ were computed via numerically differentiating the conditional likelihoods $l_t(\theta)$, $t = 1, \dots, 132$, which are a by-product of the Kalman filter employed by LS.

[10]In contrast to LS, we impose a determinate monetary policy regime throughout. Correspondingly, we adopt Del Negro and Schorfheide's (2004) prior on ψ_1 with little mass on the indeterminacy region.

[11]In a similar context, Christiano (2007) found overwhelming evidence against Gaussianity of DSGE shocks.

An alternative to sandwich posterior inference would be to directly model ε_t as, say, i.i.d. mixtures of normals. But such an approach has the same drawback as the non-Gaussian modelling of regression errors discussed in Section 5.1: The pseudo-true parameter in such a mixture specification is no longer a function of the second-order properties of Y_t, so that m.d.s.-type dependence in ε_t may well lead to an estimator of θ that is no longer consistent for the value that generates the second-order properties of Y_t.

7. CONCLUSION

In misspecified parametric models, the shape of the likelihood is asymptotically Gaussian and centered at the MLE, but of a different variance than the asymptotically normal sampling distribution of the MLE. We show that posterior beliefs constructed from such a misspecified likelihood are unreasonable in the sense that they lead to inadmissible decisions about pseudo-true values in general. Asymptotically uniformly lower risk decisions are obtained by replacing the original posterior by an artificial Gaussian posterior centered at the MLE with the sandwich covariance matrix. The sandwich covariance matrix correction, which is routinely applied for the construction of confidence regions in frequentist analyses, thus has a potentially constructive role also in Bayesian studies of potentially misspecified models.

APPENDIX

The following lemma is used in the proof of Theorem 1.

LEMMA 1: *If Σ_n, $n \geq 0$ is a sequence of stochastic matrices that are almost surely positive definite and $\Sigma_n \to \Sigma_0$ almost surely* (*in probability*), *then* $\int |\phi_{\Sigma_n}(u) - \phi_{\Sigma_0}(u)|\,du \to 0$ *almost surely* (*in probability*).

PROOF: The almost sure version follows from Problem 1 of page 132 of Dudley (2002). The convergence in probability version follows by considering almost surely converging subsequences (cf. Theorem 9.2.1 of Dudley (2002)). *Q.E.D.*

PROOF OF THEOREM 1:

(i) For any d_n, define $r_n^i(\theta_0, d_n) = E[\ell^i(\theta_0, d_n(X_n))]$, where here and below, the expectation is taken relative to P_{n,θ_0}. Note that $|r_n^i(\theta_0, d_n) - r_n(\theta_0, d_n)| \leq \sup_{a\in\mathcal{A}} |\ell_n(\theta_0, a) - \ell_n^i(\theta_0, a)| \to 0$ by Condition 4(i), so it suffices to show the claim for $r_n^i(\theta_0, d_n)$. Similarly to the notation of Section 4.3, define $\tilde{\ell}_n(u, a) = \ell_n(\theta_0 + u/\sqrt{n}, q_n(a, \theta_0))$, $\tilde{\ell}_n^i(u, a) = \ell_n^i(u/\sqrt{n}, a) = \ell_n^i(\theta_0 + u/\sqrt{n}, q_n(a, \theta_0))$, $\hat{u}_n = \sqrt{n}(\hat{\theta} - \theta_0)$, $\Sigma_{S0} = \Sigma_S(\theta_0)$, $\Sigma_{M0} = \Sigma_M(\theta_0)$, and $\tilde{\Pi}_n$ the scaled and centered posterior probability measure such that $\tilde{\Pi}_n(A) = \Pi_n(\{\theta : n^{-1/2}(\theta - \hat{\theta}) \in A\})$ for all Borel subsets $A \subset \mathbb{R}^k$. By Condition 1(ii), $\hat{\delta}_n = d_{\mathrm{TV}}(\tilde{\Pi}_n, \mathcal{N}(0, \Sigma_M)) \xrightarrow{p} 0$. Note that $\tilde{\Pi}_n$ is random measure, a probability kernel from the Borel sigma field of $\mathbb{R}^{r\times n}$ to the Borel sigma field of $\mathbb{R}^k$, indexed by the random element $X_n = D_n(\omega, \theta_0)$, $D_n : \Omega \times \Theta \mapsto \mathbb{R}^{r\times n}$.

The proof follows the logic outlined in Section 4.3. To reduce the computation of asymptotic risk to properties of nonstochastic sequences of actions, and also to deal with the stochastic nature of Σ_{M0} and Σ_{S0}, we begin by constructing an almost sure representation of the weak convergences in Condition 1. Consider first the claim about d_{Mn}.

Since $(\hat{\delta}_n, \hat{u}_n, Z, \Sigma_{S0}, \Sigma_{M0}) \Rightarrow (0, \Sigma_{S0}^{1/2} Z, Z, \Sigma_{S0}, \Sigma_{M0})$, by the Skorohod almost sure representation theorem (cf. Theorem 11.7.2 of Dudley (2002)), there exists a probability space $(\Omega^*, \mathfrak{F}^*, P^*)$ and associated random elements $(\hat{\delta}_n^*, \hat{u}_n^*, Z_n^*, \Sigma_{S0n}^*, \Sigma_{M0n}^*)$, $n \geq 1$ and $(Z^*, \Sigma_{S0}^*, \Sigma_{M0}^*)$ such that (i) $(\hat{\delta}_n^*, \hat{u}_n^*, Z_n^*, \Sigma_{S0n}^*, \Sigma_{M0n}^*) \sim (\hat{\delta}_n, \hat{u}_n, Z, \Sigma_{S0}, \Sigma_{M0})$ for all $n \geq 1$ and (ii) $(\hat{\delta}_n^*, \hat{u}_n^*, Z_n^*, \Sigma_{S0n}^*, \Sigma_{M0n}^*) \to (0, (\Sigma_{S0}^*)^{1/2} Z^*, Z^*, \Sigma_{S0}^*, \Sigma_{M0}^*)$ P^*-almost surely. Furthermore, because $\mathbb{R}^{n\times r}$ is a Polish space, by Proposition 10.2.8 of Dudley (2002), the conditional distribution of X_n given $(\hat{\delta}_n, \hat{u}_n, Z, \Sigma_{S0}, \Sigma_{M0})$ exists, for all n. Now using this conditional distribution, we can construct from $(\Omega^*, \mathfrak{F}^*, P^*)$ a probability space $(\Omega^+, \mathfrak{F}^+, P^+)$ with associated random elements $(\hat{\delta}_n^+, \hat{u}_n^+, Z_n^+, \Sigma_{S0n}^+, \Sigma_{M0n}^+, X_n^+)$, $n \geq 1$ and $(Z^+, \Sigma_{S0}^+, \Sigma_{M0}^+)$ such that (i) $(\hat{\delta}_n^+, \hat{u}_n^+, Z_n^+, \Sigma_{S0n}^+, \Sigma_{M0n}^+, X_n^+) \sim (\hat{\delta}_n, \hat{u}_n, Z, \Sigma_{S0}, \Sigma_{M0}, X_n)$ for all n and (ii) $(\hat{\delta}_n^+, \hat{u}_n^+, Z_n^+, \Sigma_{S0n}^+, \Sigma_{M0n}^+) \to (0, (\Sigma_{S0}^+)^{1/2} Z^+, Z^+, \Sigma_{S0}^+, \Sigma_{M0}^+)$ P^+-almost surely. Denote by $\tilde{\Pi}_n^+$ the posterior distribution induced by X_n^+, and write E^+ for expectations relative to P^+.

Now by definition (17), the definition of $\tilde{\ell}_n$, and $(\hat{u}_n^+, X_n^+) \sim (\hat{u}_n, X_n)$,

$$
\begin{aligned}
(37)\qquad & \inf_{a\in\mathcal{A}} \int \tilde{\ell}_n\big(u + \hat{u}_n^+, a\big)\, d\tilde{\Pi}_n^+(u) \\
& \quad = \int \tilde{\ell}_n\big(u + \hat{u}_n^+, q_n\big(d_{Mn}\big(X_n^+\big), -\theta_0\big)\big)\, d\tilde{\Pi}_n^+(u)
\end{aligned}
$$

P^+-almost surely. Also, by Condition 4(ii), $\int \tilde{\ell}_n^i(u, a_n^*(\Sigma_{M0}^+))\phi_{\Sigma_{M0}^+}(u)\,du \leq \int \tilde{\ell}_n^i(u, \hat{a}_n(X_n^+))\phi_{\Sigma_{M0}^+}(u)\,du = \int \tilde{\ell}_n^i(u + \hat{u}_n^+, q_n(\hat{a}_n(X_n^+), \hat{u}_n^+/\sqrt{n}))\phi_{\Sigma_{M0}^+}(u)\,du$ for $\hat{a}_n(X_n^+) = q_n(d_{Mn}(X_n^+), -\theta_0 - \hat{u}_n^+/\sqrt{n})$ almost surely for large enough n. Thus, similarly to (26),

$$
\begin{aligned}
(38)\qquad 0 &\leq \int \tilde{\ell}_n^i\big(u, \hat{a}_n\big(X_n^+\big)\big)\phi_{\Sigma_{M0}^+}(u)\,du - \int \tilde{\ell}_n^i\big(u, a_n^*\big(\Sigma_{M0}^+\big)\big)\phi_{\Sigma_{M0}^+}(u)\,du \\
&\leq \int \Big(\tilde{\ell}_n^i\big(u, \hat{a}_n\big(X_n^+\big)\big) - \tilde{\ell}_n\big(u + \hat{u}_n^+, q_n\big(d_{Mn}\big(X_n^+\big), -\theta_0\big)\big)\Big)\phi_{\Sigma_{M0}^+}(u)\,du \\
&\quad - \int \Big(\tilde{\ell}_n^i\big(u, a_n^*\big(\Sigma_{M0}^+\big)\big)
\end{aligned}
$$

$$
\begin{aligned}
&- \tilde{\ell}_n\big(u + \hat{u}_n^+, q_n\big(a_n^*\big(\Sigma_{M0}^+\big), \hat{u}_n^+/\sqrt{n}\big)\big)\Big)\phi_{\Sigma_{M0}^+}(u)\,du \\
&+ \int \tilde{\ell}_n\big(u + \hat{u}_n^+, q_n\big(d_{Mn}\big(X_n^+\big), -\theta_0\big)\big)\big(\phi_{\Sigma_{M0}^+}(u)\,du - d\tilde{\Pi}_n^+(u)\big) \\
&- \int \tilde{\ell}_n\big(u + \hat{u}_n^+, q_n\big(a_n^*\big(\Sigma_{M0}^+\big), \hat{u}_n^+/\sqrt{n}\big)\big)\big(\phi_{\Sigma_{M0}^+}(u)\,du - d\tilde{\Pi}_n^+(u)\big),
\end{aligned}
$$

where the inequalities hold, for each n, P^+-almost surely, so they also hold for all $n \geq 1$ P^+-almost surely. Furthermore, for any sequence $a_n \in \mathcal{A}$, by Condition 2,

$$
\begin{aligned}
&\left|\int \tilde{\ell}_n\big(u + \hat{u}_n^+, a_n\big)\big(\phi_{\Sigma_{M0}^+}(u)\,du - d\tilde{\Pi}_n^+(u)\big)\right| \\
&\quad \leq \bar{\ell} d_{\mathrm{TV}}\big(\tilde{\Pi}_n^+, \mathcal{N}\big(0, \Sigma_{M0}^+\big)\big) \\
&\quad \leq \bar{\ell}\hat{\delta}_n^+ + \bar{\ell} d_{\mathrm{TV}}\big(\mathcal{N}\big(0, \Sigma_{M0n}^+\big), \mathcal{N}\big(0, \Sigma_{M0}^+\big)\big) \to 0
\end{aligned}
$$

P^+-almost surely, since $\hat{\delta}_n^+ = d_{\mathrm{TV}}(\tilde{\Pi}_n^+, \mathcal{N}(0, \Sigma_{M0n}^+))$ and $d_{\mathrm{TV}}(\mathcal{N}(0, \Sigma_{M0n}^+), \mathcal{N}(0, \Sigma_{M0}^+)) \to 0$ P^+-almost surely by Lemma 1. Also,

$$
\begin{aligned}
&\int \big(\tilde{\ell}_n^i\big(u, q_n\big(a_n, -\hat{u}_n^+/\sqrt{n}\big)\big) - \tilde{\ell}_n\big(u + \hat{u}_n^+, a_n\big)\big)\phi_{\Sigma_{M0}^+}(u)\,du \\
&\quad = \int \big(\tilde{\ell}_n^i\big(u + \hat{u}_n^+, a_n\big) - \tilde{\ell}_n\big(u + \hat{u}_n^+, a_n\big)\big)\phi_{\Sigma_{M0}^+}(u)\,du \to 0
\end{aligned}
$$

P^+-almost surely by dominated convergence using Conditions 2 and 4(i). Thus, for P^+-almost all $\omega^+ \in \Omega^+$, the upper bound in (38) converges to zero, so that also

$$
\begin{aligned}
&\int \tilde{\ell}_n^i\big(u, \hat{a}_n\big(X_n^+\big(\omega^+\big)\big)\big)\phi_{\Sigma_{M0}^+(\omega^+)}(u)\,du \\
&\quad - \int \tilde{\ell}_n^i\big(u, a_n^*\big(\Sigma_{M0}^+\big(\omega^+\big)\big)\big)\phi_{\Sigma_{M0}^+(\omega^+)}(u)\,du \to 0
\end{aligned}
$$

and $\hat{u}_n^+(\omega^+) \to \Sigma_{S0}^+(\omega^+)^{1/2} Z^+(\omega^+)$ by construction of $(\Omega^+, \mathfrak{F}^+, P^+)$. Condition 4(iii) therefore implies that also

$$
\begin{aligned}
&\tilde{\ell}_n^i\big(-\hat{u}_n^+\big(\omega^+\big), \hat{a}_n\big(X_n^+\big(\omega^+\big)\big)\big) \\
&\quad - \tilde{\ell}_n^i\big(-\Sigma_{S0}^+\big(\omega^+\big)^{1/2} Z^+\big(\omega^+\big), a_n^*\big(\Sigma_{M0}^+\big(\omega^+\big)\big)\big) \to 0
\end{aligned}
$$

for P^+-almost all $\omega^+ \in \Omega^+$. As almost sure convergence and $\tilde{\ell}_n^i \leq \bar{\ell}$ imply convergence in expectation and $(\Sigma_{S0}^+, \Sigma_{M0}^+) \sim (\Sigma_{S0}, \Sigma_{M0})$ is independent of $Z^+ \sim \mathcal{N}(0, I_k)$, we obtain $E^+[\tilde{\ell}_n^i(-\hat{u}_n^+, \hat{a}_n(X_n^+))] - E[\int \tilde{\ell}_n^i(u, a_n^*(\Sigma_{M0})) \times$

$\phi_{\Sigma_{S0}}(u)\,du] \to 0$. But this implies, via $r_n^i(\theta_0, d_{Mn}(X_n)) = E[\tilde{\ell}_n^i(-\hat{u}_n, q_n(d_{Mn}(X_n), -\theta_0 - \hat{u}_n/\sqrt{n}))] = E^+[\tilde{\ell}_n^i(-\hat{u}_n^+, \hat{a}_n(X_n^+))]$, that also $r_n^i(\theta_0, d_{Mn}(X_n)) - E[\int \tilde{\ell}_n^i(u, a_n^*(\Sigma_{M0}))\phi_{\Sigma_{S0}}(u)\,du] \to 0$, as was to be shown.

The claim about d_{Sn} follows analogously after noting that $\int |\phi_{\hat{\Sigma}_S}(u) - \phi_{\Sigma_S(\theta_0)}(u)|\,du \xrightarrow{p} 0$ by Lemma 1.

(ii) We again focus first on the proof of the first claim. For any $\varepsilon_\eta > 0$, one can construct a continuous Lebesgue density $\dot{\eta}$ with $\int |\eta - \dot{\eta}|\,d\mu_L < \varepsilon_\eta$ that is bounded away from zero and infinity and whose compact support is a subset of the support of η—this follows from straightforward arguments after invoking, say, Corollary 1.19 of Lieb and Loss (2001). Since $|R_n(\eta, d_n) - R_n(\dot{\eta}, d_n)| < \bar{\ell}\varepsilon_\eta$, it suffices to show the claim for $R_n(\dot{\eta}, d_{Mn})$.

Pick a θ_0 in the support of $\dot{\eta}$ for which Condition 1 holds. Proceed as in the proof of part (i) and construct the random elements $(\hat{\delta}_n^*, \hat{u}_n^*, Z_n^*, \Sigma_{S0n}^*, \Sigma_{M0n}^*)$ on the probability space $(\Omega^*, \mathfrak{F}^*, P^*)$. Since the stochastic processes $\Sigma_S(\cdot)$ and $\Sigma_M(\cdot)$ may be viewed as random elements in the Polish space of continuous $\mathbb{R}^{k\times k}$ valued functions on the support of $\dot{\eta}$, the conditional distribution of $(\Sigma_S(\cdot), \Sigma_M(\cdot))$ given $(\Sigma_{S0}, \Sigma_{M0})$ exists by Proposition 10.2.8 of Dudley (2002). Further proceeding as in the proof of part (i), one can thus construct a probability space $(\Omega^+, \mathfrak{F}^+, P^+)$ with associated random elements $(\hat{\delta}_n^+, \hat{u}_n^+, Z_n^+, \Sigma_{S0n}^+, \Sigma_{M0n}^+, X_n^+)$, $n \geq 1$ and $(Z^+, \Sigma_{S0}^+, \Sigma_{M0}^+, \Sigma_S^+(\cdot), \Sigma_M^+(\cdot))$ such that (i) $(\hat{\delta}_n^+, \hat{u}_n^+, Z_n^+, \Sigma_{S0n}^+, \Sigma_{M0n}^+, X_n^+) \sim (\hat{\delta}_n, \hat{u}_n, Z, \Sigma_{S0}, \Sigma_{M0}, X_n)$ for all $n \geq 1$, $(\Sigma_{S0}^+, \Sigma_{M0}^+, \Sigma_S^+(\cdot), \Sigma_M^+(\cdot)) \sim (\Sigma_S(\theta_0), \Sigma_M(\theta_0), \Sigma_S(\cdot), \Sigma_M(\cdot))$ and $Z^+ \sim \mathcal{N}(0, I_k)$ is independent of $(\Sigma_{S0}^+, \Sigma_{M0}^+, \Sigma_S^+(\cdot), \Sigma_M^+(\cdot))$ and (ii) $(\hat{\delta}_n^+, \hat{u}_n^+, Z_n^+, \Sigma_{S0n}^+, \Sigma_{M0n}^+) \to (0, (\Sigma_{S0}^+)^{1/2}Z^+, Z^+, \Sigma_{S0}^+, \Sigma_{M0}^+)$ P^+-almost surely. Finally, for values of $\theta \in \mathbb{R}^k$ outside the support of $\dot{\eta}$, define $\Sigma_J(\theta)$ and $\Sigma_J^+(\theta)$, $J = S, M$ to equal some element of $\mathcal{P}^k$ in the support of $\Sigma_J(\theta_0)$.

Now, similarly to the proof of part (i), define

$$\delta_\phi = \int \ell_n\big(\theta_0 + \big(u + \hat{u}_n^+\big)/\sqrt{n}, d_{Mn}\big(X_n^+\big)\big)\phi_{\Sigma_M^+(\hat{\theta}_n^+)}(u)\,du - \int \ell_n\big(\theta_0 + \big(u + \hat{u}_n^+\big)/\sqrt{n}, d_{Mn}^*\big(\hat{\theta}_n^+, \Sigma_M^+\big(\hat{\theta}_n^+\big)\big)\big)\phi_{\Sigma_M^+(\hat{\theta}_n^+)}(u)\,du,$$

where $\hat{\theta}_n^+ = \theta_0 + \hat{u}_n^+/\sqrt{n}$. By Condition 5(ii), $\delta_\phi \geq 0$. Using (17), we obtain the additional inequality

$$\delta_\phi \leq \int \ell_n\big(\theta_0 + \big(u + \hat{u}_n^+\big)/\sqrt{n}, d_{Mn}\big(X_n^+\big)\big)\big(\phi_{\Sigma_M^+(\hat{\theta}_n^+)}(u)\,du - d\tilde{\Pi}_n^+(u)\big) + \int \ell_n\big(\theta_0 + \big(u + \hat{u}_n^+\big)/\sqrt{n}, d_{Mn}^*\big(\hat{\theta}_n^+, \Sigma_M^+\big(\hat{\theta}_n^+\big)\big)\big) \times \big(d\tilde{\Pi}_n^+(u) - \phi_{\Sigma_M^+(\hat{\theta}_n^+)}(u)\,du\big) \to 0,$$

and the P^+-almost sure convergence follows from $d_{\mathrm{TV}}(\mathcal{N}(0,\Sigma_{M0n}^+),\mathcal{N}(0,\Sigma_{M0}^+))\to 0$ P^+-almost surely via Lemma 1 as $\Sigma_{M0n}^+\to\Sigma_{M0}^+$ P^+-almost surely, and $\hat{\delta}_n^+=d_{\mathrm{TV}}(\tilde{\Pi}_n^+,\mathcal{N}(0,\Sigma_{M0n}^+))\to 0$ P^+-almost surely by construction. Thus, $\delta_\phi\to 0$ P^+-almost surely, too, and since $\hat{u}_n^+\to(\Sigma_{S0}^+)^{1/2}Z^+$ P^+-almost surely by construction, Condition 5(iii) yields $\ell_n(\theta_0,d_{Mn}(X_n^+))-\ell_n(\theta_0,d_n^*(\theta_0+(\Sigma_{S0}^+)^{1/2}Z^+/\sqrt{n},\Sigma_J(\theta_0+(\Sigma_{S0}^+)^{1/2}Z^+/\sqrt{n})))\to 0$ P^+-almost surely. Also, $Z^+\sim\mathcal{N}(0,I_k)$ is independent of $(\Sigma_{S0}^+,\Sigma_{M0}^+,\Sigma_S^+(\cdot),\Sigma_M^+(\cdot))\sim(\Sigma_S(\theta_0),\Sigma_M(\theta_0),\Sigma_S(\cdot),\Sigma_M(\cdot))$ and $X_n^+\sim X_n$, so that dominated convergence implies

$$r_n(\theta_0,d_{Mn})-E\biggl[\int\ell_n\bigl(\theta_0,d_{Mn}^*\bigl(\theta_0+u/\sqrt{n},\Sigma_M(\theta_0+u/\sqrt{n})\bigr)\bigr)\times\phi_{\Sigma_S(\theta_0)}(u)\,du\biggr]\to 0. \tag{39}$$

This argument can be invoked for $\dot{\eta}$-almost all θ_0, so (39) holds for $\dot{\eta}$-almost all θ_0.

Pick a large $K>0$, and define $\mathcal{B}=\{\theta\in\mathbb{R}^k:\|\Sigma_S(\theta)\|<K$ and $\|\Sigma_S(\theta)^{-1}\|<K\}$, where $\|\cdot\|$ is the spectral norm, $\dot{\ell}_n(\theta,a)=\mathbf{1}[\theta\in\mathcal{B}]\ell_n(\theta,a)$ and $\dot{r}_n(\theta,d_n)=E_\theta[\dot{\ell}_n(\theta,d_n)]$. Then

$$\dot{R}_n(\dot{\eta},d_n)=\int\dot{r}_n(\theta_0,d_n)\dot{\eta}(\theta_0)\,d\theta_0=R_n(\dot{\eta},d_n)+\varepsilon(K),$$

where $\varepsilon(K)\to 0$ as $K\to\infty$ by monotone convergence. It therefore suffices to show the claim for $\dot{R}_n(\dot{\eta},d_{Mn})$.

From (39), dominated convergence, Fubini's theorem, and a change of variables,

$$\begin{aligned}
&\int\dot{r}_n(\theta_0,d_{Mn})\dot{\eta}(\theta_0)\,d\theta_0\\
&\quad=E\int\int\dot{\ell}_n\bigl(\theta_0,d_{Mn}^*\bigl(\theta_0+u/\sqrt{n},\Sigma_M(\theta_0+u/\sqrt{n})\bigr)\bigr)\\
&\qquad\times\phi_{\Sigma_S(\theta_0)}(u)\,du\,\dot{\eta}(\theta_0)\,d\theta_0+o(1)\\
&\quad=E\int\int\dot{\ell}_n\bigl(\theta+u/\sqrt{n},d_{Mn}^*\bigl(\theta,\Sigma_M(\theta)\bigr)\bigr)\\
&\qquad\times\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\dot{\eta}(\theta+u/\sqrt{n})\,du\,d\theta+o(1).
\end{aligned}\tag{40}$$

Now consider a realization of $(\Sigma_M(\cdot),\Sigma_S(\cdot))$. Pick $\theta\in\mathcal{B}$ inside the support of $\dot{\eta}$, and define $\dot{\phi}_{\Sigma_S(t)}(u)=\mathbf{1}[t\in\mathcal{B}]\phi_{\Sigma_S(t)}(u)$. For $K_2>0$,

$$\begin{aligned}
&\int_{\|u\|\le K_2}\dot{\phi}_{\Sigma_S(\theta+u/\sqrt{n})}(u)\,du\\
&\quad\ge(2\pi)^{-k/2}\int_{\|u\|\le K_2}\mathbf{1}[\theta+u/\sqrt{n}\in\mathcal{B}]\Bigl[\inf_{\|v\|\le K_2}\det\bigl(\Sigma_S(\theta+v/\sqrt{n})\bigr)^{-1/2}\Bigr]
\end{aligned}$$

$$\cdot \exp\left[-\frac{1}{2}\sup_{\|v\|\le K_2} u'\Sigma_S(\theta+v/\sqrt{n})^{-1}u\right]du$$

$$\to \int_{\|u\|\le K_2}\phi_{\Sigma_S(\theta)}(u)\,du$$

by monotone convergence. Note that $\int_{\|u\|\le K_2}\phi_{\Sigma_S(\theta)}(u)\,du \to 1$ as $K_2\to\infty$. Also

$$\int_{\|u\|>K_2}\dot\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\,du \le (2\pi)^{-k/2}\int_{\|u\|>K_2}K^{k/2}\exp\left[-\frac{1}{2}\|u\|^2K^{-1}\right]du,$$

which is arbitrarily small for large enough K_2. Thus, $\int \dot\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\,du\to 1$, and from $\dot\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\to\phi_{\Sigma_S(\theta)}(u)$, also $\int|\dot\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)-\phi_{\Sigma_S(\theta)}(u)|\,du\to 0$ (see Problem 1 of page 132 of Dudley (2002)). Define $\dot\rho_n:\mathbb{R}^k\times\mathbb{R}^k\mapsto\mathbb{R}$ as $\dot\rho_n(\theta,u)=\dot\eta(\theta+u/\sqrt{n})/\dot\eta(\theta)$ for θ in the support of $\dot\eta$, and $\dot\rho_n(\theta,u)=0$ otherwise. Note that $\bar\rho=\sup_{\theta,u,n}\dot\rho_n(\theta,u)<\infty$ and $\dot\rho_n(\theta,u)\to 1$ by construction of $\dot\eta$, so that $\int|\dot\rho_n(\theta,u)-1|\phi_{\Sigma_S(\theta)}(u)\,du\to 0$ by dominated convergence. Therefore, $\int|\dot\rho_n(\theta,u)\dot\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)-\phi_{\Sigma_S(\theta)}(u)|\,du\le\bar\rho\int|\dot\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)-\phi_{\Sigma_S(\theta)}(u)|\,du+\int|\dot\rho_n(\theta,u)-1|\phi_{\Sigma_S(\theta)}(u)\,du\to 0$, and thus

$$\int \dot\ell_n\big(\theta+u/\sqrt{n},d^*_{Mn}\big(\theta,\Sigma_M(\theta)\big)\big)\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\dot\rho_n(\theta,u)\,du$$
$$-\int \dot\ell_n\big(\theta+u/\sqrt{n},d^*_{Mn}\big(\theta,\Sigma_M(\theta)\big)\big)\phi_{\Sigma_S(\theta)}(u)\,du\to 0.$$

This convergence holds for $\dot\eta$-almost all θ, and $\dot\rho_n$ and $\dot\ell_n$ are bounded, so dominated convergence implies

$$\int\int \dot\ell_n\big(\theta+u/\sqrt{n},d^*_{Mn}\big(\theta,\Sigma_M(\theta)\big)\big)\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\dot\eta(\theta+u/\sqrt{n})\,du\,d\theta$$
$$-\int\int \dot\ell_n\big(\theta+u/\sqrt{n},d^*_{Mn}\big(\theta,\Sigma_M(\theta)\big)\big)\phi_{\Sigma_S(\theta)}(u)\,du\,\dot\eta(\theta)\,d\theta\to 0. \tag{41}$$

Since (40) holds for almost all $(\Sigma_M(\cdot),\Sigma_S(\cdot))$, and the second term in (41) as well as (40) are bounded, it also holds in expectation, and the result follows.

The second claim follows analogously, using $d_{\mathrm{TV}}(\mathcal{N}(0,\Sigma_S(\hat\theta)),\mathcal{N}(0,\hat\Sigma_S)))\stackrel{p}{\to}0$ under P_{n,θ_0} for η-almost θ_0 from Condition 1(i), the almost sure continuity of $\Sigma_S(\cdot)$ of Condition 5(i), and Lemma 1. *Q.E.D.*

PROOF OF THEOREM 2: By straightforward arguments, assumption (iv) implies that the maximum likelihood estimator $\hat\theta=\hat\theta^m$ is consistent, $\hat\theta^m\stackrel{p}{\to}\theta_0$. Thus, there exists a real sequence $k'_n\to 0$ such that $E\mathcal{T}_n\ge 1-k'_n$, where

$\mathcal{T}_n = \mathbf{1}[\|\hat{\theta}^m - \theta_0\| < k_n']$. From now on, assume n is large enough so that $\{\theta : \|\theta - \theta_0\| < k_n'\} \subset \Theta_0$. By condition (ii) and a Taylor expansion,

$$
\begin{aligned}
0 &= \mathcal{T}_n n^{-1/2} S_n(\hat{\theta}^m) \\
&= \mathcal{T}_n n^{-1/2} S_n(\theta_0) \\
&\quad + \mathcal{T}_n \left(n^{-1} \int_0^1 H_n\bigl(\theta_0 + \lambda(\hat{\theta}^m - \theta_0)\bigr)\, d\lambda \right) n^{1/2}(\hat{\theta}^m - \theta_0)
\end{aligned}
$$

almost surely, where $H_n(\theta) = \sum_{t=1}^n h_t(\theta)$, and derivatives of the log-likelihood outside Θ_0 are defined to be zero. By assumption (v), $\mathcal{T}_n n^{-1} \| \int_0^1 H_n(\theta_0 + \lambda(\hat{\theta}^m - \theta_0))\, d\lambda - H_n(\theta_0)\| \leq \sup_{\|\theta-\theta_0\|<k_n'} n^{-1} \sum_{t=1}^n \|h_t(\theta) - h_t(\theta_0)\| \xrightarrow{p} 0$ and $n^{-1} H_n(\theta) \xrightarrow{p} -\Sigma_M^{-1}(\theta_0) = -\Sigma_{M0}^{-1}$, so that $E\mathcal{T}_n \to 1$ implies

$$
n^{1/2}(\hat{\theta}^m - \theta_0) = -\Sigma_{M0}^{-1} n^{-1/2} S_n(\theta_0) + o_p(1). \tag{42}
$$

The weak convergence in Condition 1(i) for $\hat{\theta} = \hat{\theta}^m$ now follows from (42), assumption (iii), and the continuous mapping theorem. The convergence $n^{-1} H_n(\hat{\theta}^m) \xrightarrow{p} -\Sigma_M(\theta_0)^{-1}$ follows immediately from this result and assumption (v). Furthermore, from

$$
\mathcal{T}_n s_t(\hat{\theta}^m) = \mathcal{T}_n s_t(\theta_0) + \mathcal{T}_n \left(\int_0^1 h_t\bigl(\theta_0 + \lambda(\hat{\theta}^m - \theta_0)\bigr)\, d\lambda \right) (\hat{\theta}^m - \theta_0)
$$

for $t = 1, \ldots, n$, we find

$$
\begin{aligned}
&\mathcal{T}_n \left\| n^{-1} \sum_{t=1}^n s_t(\hat{\theta}^m) s_t(\hat{\theta}^m)' - n^{-1} \sum_{t=1}^n s_t(\theta_0) s_t(\theta_0)' \right\| \\
&\quad \leq \left(\sup_{\|\theta-\theta_0\|<k_n'} n^{-1} \sum_{t=1}^n \|h_t(\theta)\| \right) \\
&\qquad \cdot \left(2\mathcal{T}_n n^{1/2} \|\hat{\theta}^m - \theta_0\| \cdot \left(\sup_{t \leq n} n^{-1/2} \|s_t(\theta_0)\| \right) \right. \\
&\qquad \left. + \mathcal{T}_n n \|\hat{\theta}^m - \theta_0\|^2 \cdot \sup_{\|\theta-\theta_0\|<k_n', t \leq n} n^{-1} \|h_t(\theta)\| \right),
\end{aligned}
$$

and $n^{-1} \sum_{t=1}^n s_t(\hat{\theta}^m) s_t(\hat{\theta}^m)' \xrightarrow{p} V(\theta_0)$ follows from the previously established $n^{1/2} \|\hat{\theta}^m - \theta_0\| = O_p(1)$ and assumptions (iii) and (v).

Define $\hat{u} = n^{1/2}(\hat{\theta}^m - \theta)$, $\hat{p} = p(\theta_0)$, $\mathrm{LR}_n(u) = \exp[L_n(\theta_0 + n^{-1/2}u) - L_n(\theta_0)]$, and $\widehat{\mathrm{LR}}_n(u) = \exp[-\frac{1}{2}u'\Sigma_{M0}^{-1}u + \hat{u}'\Sigma_{M0}^{-1}u]$. Then

$$\begin{aligned}
&d_{\mathrm{TV}}\big(\Pi_n, \mathcal{N}\big(\hat{\theta}^m, \Sigma_{M0}/n\big)\big) \\
&\quad = \int \left| \frac{p(\theta_0 + n^{-1/2}u)\mathrm{LR}_n(u)}{a_n} - \frac{\hat{p}\widehat{\mathrm{LR}}_n(u)}{\hat{a}_n} \right| du \\
&\quad \le \hat{a}_n^{-1} \int \big| p\big(\theta_0 + n^{-1/2}u\big)\mathrm{LR}_n(u) - \hat{p}\widehat{\mathrm{LR}}_n(u)\big| du + \hat{a}_n^{-1}|a_n - \hat{a}_n|,
\end{aligned}$$

where $a_n = \int p(\theta_0 + n^{-1/2}u)\mathrm{LR}_n(u)\,du > 0$ a.s. and $\hat{a}_n = \hat{p}\int \widehat{\mathrm{LR}}_n(u)\,du > 0$ a.s. Since

$$|\hat{a}_n - a_n| \le \int \big| p\big(\theta_0 + n^{-1/2}u\big)\mathrm{LR}_n(u) - \hat{p}\widehat{\mathrm{LR}}_n(u)\big|\,du, \tag{43}$$

it suffices to show that $\int |p(\theta_0 + n^{-1/2}u)\,\mathrm{LR}_n(u) - \hat{p}\widehat{\mathrm{LR}}_n(u)|\,du \xrightarrow{p} 0$ and $\hat{a}_n^{-1} = O_p(1)$. By a direct calculation, $\hat{a}_n = \hat{p}(2\pi)^{k/2}|\Sigma_{M0}|^{1/2}\exp[\frac{1}{2}\hat{u}'\Sigma_{M0}^{-1}\hat{u}]$, so that $\hat{u} = O_p(1)$ implies $\hat{a}_n = O_p(1)$ and $\hat{a}_n^{-1} = O_p(1)$.

By assumption (iv), for any natural number $m > 0$, there exists $n^*(m)$ such that, for all $n > n^*(m)$,

$$P_{n,\theta_0}\Big(\sup_{\|\theta - \theta_0\| \ge m^{-1}} n^{-1}\big(L_n(\theta) - L_n(\theta_0)\big) < -K\big(m^{-1}\big)\Big) \ge 1 - m^{-1}.$$

For any n, let m_n be the smallest m such that simultaneously, $n > \sup_{m' \le m} n^*(m')$, $n^{1/2}K(m^{-1}) > 1$, and $n^{1/2}m^{-1} > n^{1/4}$. Note that $m_n \to \infty$, since, for any fixed m, $n^*(m+1)$ and $m+1$ are finite and $K((m+1)^{-1}) > 0$. Define $\mathcal{M}_n : \mathbb{R}^k \mapsto \mathbb{R}$ as $\mathcal{M}_n(u) = \mathbf{1}[n^{-1/2}\|u\| < m_n^{-1}]$. Now

$$\begin{aligned}
&\int \big| p\big(\theta_0 + n^{-1/2}u\big)\,\mathrm{LR}_n(u) - \hat{p}\widehat{\mathrm{LR}}_n(u)\big|\,du \\
&\quad \le \int \big| p\big(\theta_0 + n^{-1/2}u\big)\mathcal{M}_n(u)\,\mathrm{LR}_n(u) - \hat{p}\widehat{\mathrm{LR}}_n(u)\big|\,du \\
&\qquad + \int \big(1 - \mathcal{M}_n(u)\big)\,p\big(\theta_0 + n^{-1/2}u\big)\,\mathrm{LR}_n(u)\,du,
\end{aligned}$$

and by construction of $\mathcal{M}_n(u)$, with probability of at least $1 - m_n^{-1}$,

$$\begin{aligned}
&\int \big(1 - \mathcal{M}_n(u)\big)\,p\big(\theta_0 + n^{-1/2}u\big)\,\mathrm{LR}_n(u)\,du \\
&\quad \le \int p\big(\theta_0 + n^{-1/2}u\big)\,du \cdot \sup_{\|\theta - \theta_0\| \ge m_n^{-1}} \exp\big[L_n(\theta) - L_n(\theta_0)\big] \\
&\quad \le n^{k/2}\exp\big[-n \cdot K\big(m_n^{-1}\big)\big] \le n^{k/2}\exp\big[-n^{1/2}\big] \to 0.
\end{aligned}$$

Furthermore, with $\zeta_n = \int |\mathcal{M}_n(u)\mathrm{LR}_n(u) - \widehat{\mathrm{LR}}_n(u)|\, du$,

$$\begin{aligned}&\int \big|p\big(\theta_0 + n^{-1/2}u\big)\mathcal{M}_n(u)\,\mathrm{LR}_n(u) - \hat{p}\widehat{\mathrm{LR}}_n(u)\big|\, du \\ &\quad \le \int \big|p\big(\theta_0 + n^{-1/2}u\big) - \hat{p}\big|\mathcal{M}_n(u)\,\mathrm{LR}_n(u)\, du + \hat{p}\zeta_n\end{aligned}$$

and

$$\begin{aligned}&\int \big|p\big(\theta_0 + n^{-1/2}u\big) - \hat{p}\big|\mathcal{M}_n(u)\,\mathrm{LR}_n(u)\, du \\ &\quad \le (\zeta_n + \hat{a}_n/\hat{p}) \cdot \sup_{\|\theta-\theta_0\|\le m_n^{-1}} \big|p(\theta) - \hat{p}\big|.\end{aligned}$$

By assumption (i), $p(\theta)$ is continuous at θ_0, so $\sup_{\|\theta-\theta_0\|\le m_n^{-1}} |p(\theta) - \hat{p}| \to 0$. Furthermore, $\hat{a}_n = O_p(1)$ as shown above, so it suffices to prove that $\zeta_n \xrightarrow{p} 0$ to obtain $d_{\mathrm{TV}}(\Pi_n, \mathcal{N}(\hat{\theta}^m, \Sigma_{M0}/n)) \xrightarrow{p} 0$.

By an exact Taylor expansion, for any $u \in \mathbb{R}^k$ satisfying $\theta_0 + n^{-1/2}u \in \Theta_0$,

$$\begin{aligned}&L_n\big(\theta_0 + n^{-1/2}u\big) - L_n(\theta_0) \\ &\quad = n^{-1/2}S_n(\theta_0) + \frac{1}{2}u'\left(n^{-1}\int_0^1 H_n\big(\theta_0 + \lambda n^{-1/2}u\big)\, d\lambda\right)u\end{aligned}$$

almost surely. Thus, for all n large enough to ensure $\{\theta : \|\theta - \theta_0\| < m_n^{-1}\} \subset \Theta_0$, also

$$\sup_{u\in\mathbb{R}^k} \mathcal{M}_n(u)\left|\mathrm{LR}_n(u)/\widehat{\mathrm{LR}}_n(u) - \exp\left[\delta_n' u + \frac{1}{2}u'\Delta_n(u)u\right]\right| = 0$$

almost surely, where $\delta_n = n^{-1/2}S_n(\theta_0) - \Sigma_{M0}^{-1}\hat{u}$ and $\Delta_n(u) = n^{-1}\int_0^1 H_n(\theta_0 + \lambda n^{-1/2}u)\, d\lambda + \Sigma_{M0}^{-1}$. By Jensen's inequality,

$$\begin{aligned}\zeta_n &= \hat{a}_n \int \left|1 - \mathcal{M}_n(u)\exp\left[\delta_n' u + \frac{1}{2}u'\Delta_n(u)u\right]\right| \phi_{\Sigma_{M0}}(u - \hat{u})\, du \qquad (44) \\ &\le \hat{a}_n\left(\int \left(1 - \mathcal{M}_n(u)\exp\left[\delta_n' u + \frac{1}{2}u'\Delta_n(u)u\right]\right)^2 \right. \\ &\qquad \left. \times\, \phi_{\Sigma_{M0}}(u - \hat{u})\, du\right)^{1/2}\end{aligned}$$

almost surely. By assumption (v),

$$\begin{aligned}\mathcal{M}_n(u)\|\Delta_n(u)\| &\le c_n \\ &= \sup_{\|\theta-\theta_0\|\le m_n^{-1}} n^{-1}\|H_n(\theta)-H_n(\theta_0)\| \\ &\quad + \|n^{-1}H_n(\theta_0)+\Sigma_{M0}^{-1}\| \xrightarrow{p} 0\end{aligned}$$

and

$$\begin{aligned}&\int \mathcal{M}_n(u)\exp[2\delta_n' u + u'\Delta_n(u)u]\phi_{\Sigma_{M0}}(u-\hat{u})\,du \\ &\quad\le \int \exp[2\delta_n' u + c_n u'u]\phi_{\Sigma_{M0}}(u-\hat{u})\,du, \\ &\int \mathcal{M}_n(u)\exp\left[\delta_n' u + \frac{1}{2}u'\Delta_n(u)u\right]\phi_{\Sigma_{M0}}(u-\hat{u})\,du \\ &\quad\ge \int \exp\left[\delta_n' u - \frac{1}{2}c_n u'u\right]\phi_{\Sigma_{M0}}(u-\hat{u})\,du \\ &\qquad - \int \big(1-\mathcal{M}_n(u)\big)\exp\left[\delta_n' u + \frac{1}{2}c_n u'u\right]\phi_{\Sigma_{M0}}(u-\hat{u})\,du\end{aligned}$$

almost surely. From (42), $\delta_n \xrightarrow{p} 0$, so that $\int \exp[2\delta_n' u + c_n u'u]\phi_{\Sigma_{M0}}(u-\hat{u})\,du \xrightarrow{p} 1$ and $\int \exp[\delta_n' u - \frac{1}{2}c_n u'u]\phi_{\Sigma_{M0}}(u-\hat{u})\,du \xrightarrow{p} 1$. Finally, by another application of the Cauchy–Schwarz inequality,

$$\begin{aligned}&\left(\int \big(1-\mathcal{M}_n(u)\big)\exp\left[\delta_n' u - \frac{1}{2}c_n u'u\right]\phi_{\Sigma_{M0}}(u-\hat{u})\,du\right)^2 \\ &\quad\le \int \big(1-\mathcal{M}_n(u)\big)\phi_{\Sigma_{M0}}(u-\hat{u})\,du \\ &\qquad \cdot \int \exp[2\delta_n' u + c_n u'u]\phi_{\Sigma_{M0}}(u-\hat{u})\,du \xrightarrow{p} 0,\end{aligned}$$

and the convergence follows from $\int (1-\mathcal{M}_n(u))\phi_{\Sigma_{M0}}(u-\hat{u})\,du = \int_{\|u\|\ge n^{1/2}m_n^{-1}} \phi_{\Sigma_{M0}}(u-\hat{u})\,du \xrightarrow{p} 0$ and the same arguments as above. Thus, the right-hand side of (44) converges in probability to zero, and $\zeta_n \ge 0$, so that $\zeta_n \xrightarrow{p} 0$.

Thus, $d_{\mathrm{TV}}(\Pi_n, \mathcal{N}(\hat{\theta}^m, \Sigma_{M0}/n)) \xrightarrow{p} 0$, which implies that the posterior median $\hat{\theta}^\Pi$ satisfies $n^{1/2}(\hat{\theta}^\Pi - \hat{\theta}^m) \xrightarrow{p} 0$, and $n^{-1}\sum_{t=1}^n s_t(\hat{\theta}^\Pi)s_t(\hat{\theta}^\Pi)' \xrightarrow{p} V(\theta_0)$ follows from the same arguments used for $\hat{\theta} = \hat{\theta}^m$ above. Finally, $d_{\mathrm{TV}}(\Pi_n, \mathcal{N}(\hat{\theta}^m,$

$\Sigma_{M0}/n)) \xrightarrow{p} 0$ also implies that the posterior asymptotic variance of Π_n converges in probability to Σ_{M0}. *Q.E.D.*

REFERENCES

ANDREWS, D. W. K. (1987): "Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers," *Econometrica*, 55, 1465–1471. [1827]

BAYARRI, M. J., AND J. O. BERGER (1997): "Measures of Surprise in Bayesian Analysis," Working Paper 97-46, Duke University. [1826]

BERGER, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis* (Second Ed.). New York: Springer-Verlag. [1813]

BERK, R. H. (1966): "Limiting Behavior of Posterior Distributions When the Model Is Incorrect," *Annals of Mathematical Statistics*, 37, 51–58. [1814]

——— (1970): "Consistency a posteriori," *Annals of Mathematical Statistics*, 41, 894–906. [1814]

BOOS, D. D., AND J. F. MONAHAN (1986): "Bootstrap Methods Using Prior Information," *Biometrika*, 73, 77–83. [1808]

BOX, G. E. P. (1980): "Sampling and Bayes' Inference in Scientific Modelling," *Journal of the Royal Statistical Society, Ser. A*, 143, 383–430. [1826]

BUNKE, O., AND X. MILHAUD (1998): "Asymptotic Behavior of Bayes Estimates Under Possibly Incorrect Models," *The Annals of Statistics*, 26, 617–644. [1817]

CHEN, C. (1985): "On Asymptotic Normality of Limiting Density Functions With Bayesian Implications," *Journal of the Royal Statistical Society, Ser. B*, 47, 540–546. [1817]

CHOW, G. C. (1984): "Maximum-Likelihood Estimation of Misspecified Models," *Economic Modelling*, 1, 134–138. [1817]

CHRISTIANO, L. J. (2007): "Comment," *Journal of Business & Economic Statistics*, 25, 143–151. [1836]

DEL NEGRO, M., AND F. SCHORFHEIDE (2004): "Priors From General Equilibrium Models for VARs," *International Economic Review*, 45, 643–673. [1836]

DOKSUM, K. A., AND A. Y. LO (1990): "Consistent and Robust Bayes Procedures for Location Based on Partial Information," *The Annals of Statistics*, 18, 443–453. [1808]

DOMOWITZ, I., AND H. WHITE (1982): "Misspecified Models With Dependent Observations," *Journal of Econometrics*, 20, 35–58. [1817]

DUDLEY, R. M. (2002): *Real Analysis and Probability*. Cambridge, U.K.: Cambridge University Press. [1837,1838,1840,1842]

FAHRMEIR, L., AND G. TUTZ (2001): *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag. [1814]

FREEDMAN, D. A. (2006): "On the So-Called 'Huber Sandwich Estimator' and 'Robust Standard Errors'," *The American Statistician*, 60, 299–302. [1807]

GELMAN, A., X. MENG, AND H. STERN (1996): "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," *Statistica Sinica*, 6, 733–807. [1826]

GELMAN, A., J. B. CARLIN, H. S. STERN, AND D. B. RUBIN (2004): *Bayesian Data Analysis* (Second Ed.). Boca Raton, FL: Chapman & Hall/CRC. [1807,1830]

GEWEKE, J. (2005): *Contemporary Bayesian Econometrics and Statistics*. Hoboken, NJ: Wiley. [1807]

GOURIEROUX, C., A. MONFORT, AND A. TROGNON (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700. [1814]

HALL, P., AND C. C. HEYDE (1980): *Martingale Limit Theory and Its Applications*. New York: Academic Press. [1826,1827]

HARTIGAN, J. (1983): *Bayes Theory*. New York: Springer. [1807]

HUBER, P. (1967): "The Behavior of the Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 221–233. [1806]

JEGANATHAN, P. (1995): "Some Aspects of Asymptotic Theory With Applications to Time Series Models," *Econometric Theory*, 11, 818–887. [1817]

KIM, J. (2002): "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, 107, 175–193. [1808]

KLEIJN, B. J. K., AND A. W. VAN DER VAART (2012): "The Bernstein–Von-Mises Theorem Under Misspecification," *Electronic Journal of Statistics*, 6, 354–381. [1817]

KWAN, Y. K. (1999): "Asymptotic Bayesian Analysis Based on a Limited Information Estimator," *Journal of Econometrics*, 88, 99–121. [1808]

LANCASTER, T. (2003): "A Note on Bootstraps and Robustness," Working Paper, Brown University. [1825]

LIEB, E. H., AND M. LOSS (2001): *Analysis* (Second Ed.). Providence, RI: American Mathematical Society. [1840]

LUBIK, T., AND F. SCHORFHEIDE (2004): "Testing for Indeterminacy: An Application to U.S. Monetary Policy," *American Economic Review*, 94, 190–217. [1835]

MORAN, P. A. P. (1948): "Rank Correlation and Product-Moment Correlation," *Biometrika*, 35, 203–206. [1833]

MÜLLER, U. K. (2013): "Supplement to 'Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix'," *Econometrica Supplemental Material*, 81, http://www.econometricsociety.org/ecta/supmat/9097_data_and_programs.zip. [1808]

NEWEY, W. K., AND K. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [1833]

PELENIS, J. (2010): "Bayesian Semiparametric Regression," Working Paper, Princeton University. [1830]

PRATT, J. W., H. RAIFFA, AND R. SCHLAIFER (1965): *Introduction to Statistical Decision Theory*. New York: Wiley. [1808]

ROYALL, R., AND T. TSOU (2003): "Interpreting Statistical Evidence by Using Imperfect Models: Robust Adjusted Likelihood Functions," *Journal of the Royal Statistical Society, Ser. B*, 65, 391–404. [1807,1808,1824,1825]

SCHENNACH, S. M. (2005): "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika*, 92, 31–46. [1825]

SCHERVISH, M. J. (1995): *Theory of Statistics*. New York: Springer. [1811,1827]

SCHORFHEIDE, F. (2008): "DSGE Model-Based Estimation of the New Keynesian Phillips Curve," *FRB Richmond Economic Quarterly*, 94, 397–433. [1836]

SHALIZI, C. R. (2009): "Dynamics of Bayesian Updating With Dependent Data and Misspecified Models," *Electronic Journal of Statistics*, 3, 1039–1074. [1817]

SILVERMAN, B. W. (1986): *Density Estimation*. London: Chapman & Hall. [1832]

STAFFORD, J. E. (1996): "A Robust Adjustment of the Profile Likelihood," *The Annals of Statistics*, 24, 336–352. [1808,1824]

SZPIRO, A. A., K. M. RICE, AND T. LUMLEY (2010): "Model-Robust Regression and a Bayesian 'Sandwich' Estimator," *The Annals of Applied Statistics*, 4, 2099–2113. [1825]

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge, U.K.: Cambridge University Press. [1829]

WHITE, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–830. [1825]

——— (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [1806,1817]

ZELLNER, A. (1997): "The Bayesian Method of Moments (BMOM): Theory and Applications," in *Applying Maximum Entropy to Econometric Problems*. Advances in Econometrics, Vol. 12, ed. by T. B. Fomby and R. C. Hill. Bingley: Emerald Group Publishing, 85–105. [1808]

Dept. of Economics, Princeton University, Princeton, NJ 08544, U.S.A.; umueller@princeton.edu.

Manuscript received June, 2009; final revision received August, 2012.