Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix*

Ulrich K. Müller
Princeton University
Department of Economics

First verstion: June 2009
This version: November 2009

Abstract

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 frquentist 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.

JEL classification: C44, C11

Keywords: posterior variance, quasi-likelihood, pseudo-true parameter value, interval estimation

*I thank Andriy Norets and Chris Sims, as well as seminar participants at NYU, Princeton, Rice, UBC and Virginia for helpful discussions, Jia Li for excellent research assistance, and Christopher Otrok for his help with the data underlying Section 5. Financial support by the NSF through grant SES-0751056 is gratefully acknowledged.

1 Introduction

A major attraction of Bayesian inference stems from classical decision theory: minimizing Bayesian posterior loss for each observed data set generally minimizes Bayes risk. A concern for average frequentist risk thus naturally leads to the Bayesian paradigm as the optimal mode of inference. This implication, however, requires in general a correctly specified likelihood of the data.

The seminal results of Huber (1967) and White (1982) provide the asymptotic *sampling distribution* of the maximum likelihood estimator (MLE) in misspecified models: It is concentrated on the Kullback-Leibler divergence minimizing pseudo-true value and, to first asymptotic order, it is Gaussian with the "sandwich" covariance matrix. This sandwich matrix involves both the second derivative of the log-likelihood and the variance of the scores. In a number of instances, pseudo-true parameter values remain the natural object of interest also under misspecification. Estimators of the sandwich covariance matrix are thus prevalent in frequentist applied work, as valid confidence regions must, by definition, reflect the sampling variability of the MLE.

It is also well known how the *posterior* of parametric models behaves asymptotically under misspecification: It is Gaussian, centered at the MLE and with covariance matrix equal to the inverse of the second derivative of the log-likelihood. See, for instance, Section 4.2 and Appendix B in Gelman, Carlin, Stern, and Rubin (2004) or Section 3.4 of Geweke (2005) for textbook treatments, and Blackwell (1985), Chen (1985), Bunke and Milhaud (1998) and Kleijn and van der Vaart (2008) for formal expositions. The large sample posterior thus behaves *as if* one had observed a normally distributed MLE with variance equal to the inverse of the second derivative of the log-likelihood.

But this asymptotic posterior variance does not correspond in general to the sandwich covariance matrix of the actual asymptotic sampling distribution of the MLE: There is a mismatch between the perceived accuracy of the information about the pseudo-true parameter value in the posterior and the actual accuracy as measured by the sampling distribution of the MLE. As long as loss is a function of the pseudo-true parameter value, this suggests that one obtains decisions of lower risk (that is, lower expected loss over repeated samples) by replacing the actual posterior by a Gaussian "sandwich" posterior centered at the MLE with sandwich covariance matrix. The main point of this paper is to formally analyze this intuition, and to suggest a procedure to implement the sandwich correction in models with high dimensional nuisance parameters.

The relatively closest contribution in the literature seems to be a one page discussion in Royall and Tsou (2003). They consider 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 sampling variance of the scalar (profile) MLE to first order. In their Section 8, Royall and Tsou verbally discuss 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 note 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 go on to mistakenly claim that otherwise, the posterior based on the adjusted likelihood is conservative in the sense of overstating the variance. See comment 2 in Section 3.2 below for further discussion.

It is a crucial assumption of that the pseudo-true parameter of the misspecified model remains the object of interest, as also stressed by Royall and Tsou (2003) and Freedman (2006). For instance, consider a linear regression with mean independent disturbances, and suppose that the parameter of interest is the population regression coefficient. The pseudo-true parameter value of the normal linear model remains the population coefficient for any regression with mean independent disturbances. In contrast, a linear model with, say, disturbances that are mixtures of normals independent of the regressors does not in general yield a pseudo-true value equal to the population regression coefficient. We numerically demonstrate the impact of this effect on risk in Section 4 below.

In models with a high dimensional parameter it might not be possible or reasonable to rely on the usual maximum likelihood approximations. In fact, one important practical appeal of Bayesian inference is precisely that it can handle models with high dimensional nuisance parameters, some of which might not be tightly identified by the likelihood. This raises the question of how to implement the sandwich posterior correction in such models. One possibility is to integrate out (some of) the nuisance parameters over their prior and to base inference on the resulting integrated likelihood of the parameters of interest. Formally, this integrated likelihood is the likelihood of a model where the nuisance parameters are drawn at random from their prior distribution, and the data is then drawn from the original model conditional on the realization of the nuisance parameters. This approach is a practically useful compromise between a fully fledged Bayesian analysis (which assumes correct specification of the likelihood in all respects) and standard maximum likelihood estimation

of all parameters with sandwich covariance matrix (which, if it was implementable, would allow for a pseudo-true interpretation of all parameters and would not require a prior at all). It is shown that an appropriate sequence of scores of the integrated likelihood can be obtained by computing posterior averages of the partial score of the original model conditional on increasing subsets of the observed data. Well-developed posterior sampling algorithms can thus be put to use to compute a sandwich posterior that corrects at least for some forms of misspecification.

As an empirical illustration we consider the unobserved factor model of Kose, Otrok, and Whiteman (2003), who model the co-movement of output, consumption and investment growth in a panel data set of 60 countries by a world, regional and country specific factors. We find that sandwich variances of the world factor are approximately 2 to 10 times larger than the uncorrected posterior variances. The world factor is thus considerably less precisely identified by the data than implied by an analysis that does not account for potential misspecification.

The remainder of the paper is organized as follows. Section 2 below provides a heuristic derivation of the large sample superiority of Bayesian inference based on the sandwich posterior in misspecified models, and of the suggested algorithm for the implementation of the sandwich posterior in highly dimensional models. Section 3 contains the formal discussion. The small sample results for a linear regression model are in Section 4, and Section 5 contains the empirical application to the Kose, Otrok, and Whiteman (2003)-model. Section 6 concludes.

2 Heuristics

2.1 Large Sample Approximations in a Model with Independent Observations

2.1.1 Misspecified Models and Pseudo-True Parameter Values

Let x_i, $i = 1, \cdots, n$ 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 assumed, yielding a log-likelihood equal to $L_n(\theta) = \sum_{i=1}^n \ln g(x_i, \theta)$. If $f(x) \neq g(x, \theta)$ for all $\theta \in \Theta$, then the assumed model $g(x, \theta)$ is misspecified. Let $\hat{\theta}$ be the MLE, $L_n(\hat{\theta}) = \sup_{\theta \in \Theta} L_n(\theta)$. Since $n^{-1} L_n(\theta) \xrightarrow{p} l_0(\theta) = E \ln g(x_i, \theta)$ 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. If f is absolutely continuous with respect to g, then

$$l_0(\theta) - E \ln f(x_i) = -\int f(x) \ln \frac{f(x)}{g(x,\theta)} d\mu(x) = -K(\theta) \tag{1}$$

where $K(\theta)$ is the Kullback-Leibler divergence between the models $f(x)$ and $g(x, \theta)$, so θ_0 is also the Kullback-Leibler minimizing value $\theta_0 = \arg\min_{\theta\in\Theta} K(\theta)$. In the correctly specified model, θ_0 is simply the true, data generating value. In misspecified models, this "pseudo-true" value θ_0 sometimes remains the natural object of interest. As mentioned in the introduction, the assumption of Gaussian disturbances 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 disturbance is 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 also arise in exponential models with correctly specified mean, 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)). In the following development, we follow the frequentist quasi-likelihood literature and assume that the object of interest in a misspecified model is this pseudo-true parameter value.

2.1.2 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

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

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 (2)

$$n^{1/2}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_S) \tag{3}$$

where $\Sigma_S = \Sigma_M V \Sigma_M$ and $V = E[s_i(\theta_0) s_i(\theta_0)']$. (The subscripts S and M stand for "Sandwich" and "Model", respectively, since in a correctly specified model, $\Sigma_S = \Sigma_M$ via

the information equality $V = \Sigma_M^{-1}$). Thus, by definition, asymptotically justified confidence sets for θ_0 must be based on the sandwich form $\Sigma_S = \Sigma_M V \Sigma_M$. This is typically implemented by estimating Σ_M and V by $\hat{\Sigma}_M = \left(-n^{-1}\sum_{i=1}^n h_i(\hat{\theta})\right)^{-1} \xrightarrow{p} \Sigma_M(\theta_0)$ and $\hat{V} = n^{-1}\sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})' \xrightarrow{p} V(\theta_0)$, and $\hat{\Sigma}_S = \hat{\Sigma}_M \hat{V} \hat{\Sigma}_M \xrightarrow{p} \Sigma_S(\theta_0)$. For simplicity, we develop the heuristics in Section 2 under the assumption that $\Sigma_M(\theta_0) = \Sigma_M$ and $V(\theta_0) = V$ do not depend on θ_0.

2.1.3 Large Sample Properties of the Likelihood

From a Bayesian perspective, the sample information about θ_0 is contained in the likelihood $L_n(\theta)$. From a second order Taylor expansion of $L_n(\theta)$ around $\hat{\theta}$, we obtain for all fixed $u \in \mathbb{R}^k$

$$L_n(\hat{\theta} + n^{-1/2}u) - L_n(\hat{\theta}) \xrightarrow{p} -\tfrac{1}{2}u'\Sigma_M^{-1}u \tag{4}$$

because $\sum_{i=1}^n s_i(\hat{\theta}) = 0$. This suggests that in large samples, the sample information about θ_0 conveyed by the likelihood is that of a Gaussian random variable with mean $\hat{\theta}$ and variance Σ_M/n. Bayesian posterior expected loss in the misspecified model will thus behave *as if* one had observed $\hat{\theta} \sim \mathcal{N}(\theta_0, \Sigma_M/n)$. But, as noted above, the actual sampling distribution of $\hat{\theta}$ is approximately $\hat{\theta} \sim \mathcal{N}(\theta_0, \Sigma_S/n)$, with $\Sigma_S \neq \Sigma_M$ in general. This suggests that in misspecified models, the likelihood provides a misleading account of the sample information about θ_0, and that one would do better by replacing the misspecified log-likelihood by the "sandwich" log-likelihood L_{Sn} from the model $\hat{\theta} \sim \mathcal{N}(\theta_0, \hat{\Sigma}_S/n)$,

$$L_{Sn}(\theta) = C - \tfrac{1}{2}n(\theta - \hat{\theta})'\hat{\Sigma}_S^{-1}(\theta - \hat{\theta}) \tag{5}$$

where C here and below is a generic constant.

2.2 Single Observation Gaussian Location Problem

Large sample Bayesian inference about θ based on the original likelihood (4) thus approximately behaves like Bayesian inference in the model where the single $k \times 1$ random vector Y has actual sampling distribution

$$Y \sim \mathcal{N}(\theta, \Sigma_S/n), \tag{6}$$

but it is mistakenly assumed that $Y \sim \mathcal{N}(\theta, \Sigma_M/n)$. Similarly, large sample Bayesian inference based on the sandwich likelihood (5) corresponds to inference in (6) using the

correct model $Y \sim \mathcal{N}(\theta, \Sigma_S/n)$. To compare these two approaches, let the topological space $\mathcal{A}$ be comprised of all possible actions, and let $\mathcal{D}$ be the (non-randomized) decision rules after observing (6), that is measurable mappings $\mathbb{R}^k \mapsto \mathcal{A}$. Introduce the loss function $\ell : \Theta \times \mathcal{A} \mapsto [0, \infty)$, which we assume to be non-negative. The frequentist risk of decision $d \in \mathcal{D}$ is given by

$$r(\theta, d) = E_\theta[\ell(\theta, d(Y)] = \int \ell(\theta, d(y))\phi_{\Sigma_S/n}(y - \theta)dy$$

where $\phi_{\Sigma_S/n}$ is the density of the measure $\mathcal{N}(0, \Sigma_S/n)$, and from the point of view of classical decision theory, good decisions are those that yield low risk. Let $p(\theta)$ be the prior Lebesgue probability density on Θ. The Bayes risk of decision d relative to the prior p equals

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

Note that

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

where the interchange of the order of integration is allowed by Fubini's Theorem. If $d_S^* \in \mathcal{D}$ is the decision that minimizes posterior expected loss for each observation $Y = y$, so that d_S^* satisfies

$$\int \ell(\theta, d_S^*(y))\phi_{\Sigma_S/n}(y - \theta)p(\theta)d\theta = \inf_{a \in \mathcal{A}} \int \ell(\theta, a)\phi_{\Sigma_S/n}(y - \theta)p(\theta)d\theta$$

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.

In contrast, the mistaken assumption $Y \sim \mathcal{N}(\theta, \Sigma_M/n)$, $\Sigma_M \neq \Sigma_S$, will in general lead to the decision $d_M^* \in \mathcal{D}$ that satisfies, for each y,

$$\int \ell(\theta, d_M^*(y))\phi_{\Sigma_M/n}(y - \theta)p(\theta)d\theta = \inf_{a \in \mathcal{A}} \int \ell(\theta, a)\phi_{\Sigma_M/n}(y - \theta)p(\theta)d\theta.$$

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

Example 1 *Suppose* $\Theta = \mathbb{R}$, $p(\theta) = \phi_{\sigma_p^2}(\theta - \mu_p)$, $A = \Theta$ *and* $\ell(\theta, a) = (\theta - a)^2$. *Then, for* $J = S, M$, $d_J^*(y) = (\mu_p \Sigma_J/n + \sigma_p^2 y)/(\Sigma_J/n + \sigma_p^2)$, *and* $R(p, d_S^*) < R(p, d_M^*)$.

2.2.1 Dominating Sample Information

Now suppose n is large, and $p(\theta)$ is continuous. Then, for $J = S, M$, the posterior is proportional to $\phi_{\Sigma_J/n}(y-\theta)p(\theta) \approx \phi_{\Sigma_J/n}(y-\theta)p(y)$ with the likelihood dominating the prior, so that the posterior simply becomes approximately equal to $\mathcal{N}(y, \Sigma_J/n)$. The best decision d_J^* then satisfies

$$\int \ell(\theta, d_J^*(y))\phi_{\Sigma_J/n}(y-\theta)d\theta = \inf_{a\in\mathcal{A}} \int \ell(\theta, a)\phi_{\Sigma_J/n}(y-\theta)d\theta \tag{7}$$

for each realization of y. Note that this is recognized as the best decision rule relative to the improper prior equal to Lebesgue measure. Also

$$\begin{aligned} R(p,d) &= \int\int \ell(\theta, d(y))\phi_{\Sigma_S/n}(y-\theta)p(\theta)d\theta dy \\ &\approx \int\int \ell(\theta, d(y))\phi_{\Sigma_S/n}(y-\theta)d\theta p(y)dy. \end{aligned} \tag{8}$$

Thus, as before, $R(p, d_M^*) \geq R(p, d_S^*)$. In fact, because the the optimality of d_S^* stems from the pointwise comparison of the integrand in (8), one obtains more generally $R(\eta, d_M^*) \geq R(\eta, d_S^*)$ for any continuous density η. As before, one would expect that for many decision problems, $R(\eta, d_M^*) > R(\eta, d_S^*)$, but now the the optimal decision must vary as a function of the maintained variance for the strict inequality to hold.

Example 2 *Under squared loss as in Example 1, the decision d_J^* of (7) are identical, $d_S^* = d_M^*$, so there is no difference in their risks. But if squared loss is replaced by, say, the asymmetric "linex" loss function $\ell(\theta, a) = \exp[b(\theta-a)] - b(\theta-a) - 1$, $b \neq 0$, one obtains the variance dependent optimal rule $d_J^* = y + \frac{1}{2}b\Sigma_J/n$.*

Also, consider the interval estimation problem with $\Theta = \mathbb{R}^k$, p an everywhere continuous Lebesgue probability density, $A = (a_l, a_u) \in \mathbb{R}^2$, $a_l \leq a_u$, and $\ell(\theta, a) = a_u - a_l + c(\mathbf{1}[\theta_1 < a_l](a_l - \theta_{(1)}) + \mathbf{1}[\theta > a_u](\theta_{(1)} - a_u))$, where $\theta_{(1)}$ is the first element of θ. Then it is easy to see (cf. Theorem 5.78 of Schervish (1995)) that the usual two-sided equal-tailed posterior probability interval $d_J^*(y) = [y - m_J^*, y + m_J^*]$ with $m_J^* = z_c\sqrt{\Sigma_{J(1,1)}/n}$ the $1 - c^{-1}$ quantile of $\mathcal{N}(0, \Sigma_{J(1,1)}/n)$, satisfies (7), where $\Sigma_{J(1,1)}$ is the $(1,1)$ element of Σ_J. Further, a calculation shows that $\int \ell(\theta, d_S^*(y))\phi_{\Sigma_S/n}(y-\theta)d\theta < \int \ell(\theta, d_M^*(y))\phi_{\Sigma_S/n}(y-\theta)d\theta$ for all y, so that also $R(p, d_S^*) < R(p, d_M^*)$.

2.2.2 Invariant Loss

A further simplification beyond (7) arises if ℓ is invariant, that is if for some function $q : \Theta \times \mathcal{A} \mapsto \mathcal{A}$, $\ell(\theta_1, a) = \ell(\theta_2, q(\theta_1 - \theta_2, a))$ for all $a \in \mathcal{A}$, $\theta_1, \theta_2 \in \Theta$. If a_J^*, $J = S, M$ minimizes expected loss when after observing $Y = 0$,

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

then the invariant rule $d_S^*(y) = q(y, a_J^*)$ is best under the maintained model $Y \sim \mathcal{N}(\theta, \Sigma_J/n)$, since

$$\int \ell(\theta, q(y, a))\phi_{\Sigma_J/n}(y - \theta)d\theta = \int \ell(\theta, a)\phi_{\Sigma_J/n}(-\theta)d\theta. \tag{9}$$

Furthermore, for any invariant rule $d(y) = q(y, a)$

$$\begin{aligned} r(\theta, d) &= \int \ell(\theta, q(y, a))\phi_{\Sigma_S/n}(y - \theta)dy \\ &= \int \ell(\theta - y, a)\phi_{\Sigma_S/n}(y - \theta)dy \\ &= \int \ell(y, a)\phi_{\Sigma_S/n}(-y)dy. \end{aligned} \tag{10}$$

Thus, the frequentist risk $r(\theta, d_S^*)$ of the invariant rule d_S^* is equal to its posterior expected loss with the improper Lebesgue prior (9), $\int \ell(\theta, a_S^*)\phi_{\Sigma_S/n}(-\theta)d\theta$, and 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 $\theta \in \Theta$, $r(\theta, d_S^*) \leq r(\theta, d_M^*) = \int \ell(y, a_M^*)\phi_{\Sigma_S/n}(-y)dy$, with equality only if the optimal action a_J^* does not depend on the posterior variance Σ_J/n.

Example 3 *The interval estimation loss function of Example 2 is easily seen to be invariant with* $q(\theta, a) = [a_l + \theta_{(1)}, a_u + \theta_{(1)}]$. *Thus, by (10),* $r(\theta, d_J^*) = \int \ell(y, (-m_J^*, m_J^*))\phi_{\Sigma_S/n}(-y)dy$, *and the result* $R(p, d_S^*) < R(p, d_M^*)$ *of Example 2 is strengthened to* $r(\theta, d_S^*) < r(\theta, d_M^*)$ *for* all $\theta \in \Theta$.

Another example of a set estimation problem arises with $\Theta = \mathbb{R}^k$, $A = \{$*all Borel subsets of* $\mathbb{R}^k\}$ *and* $\ell(\theta, a) = \mu_L(a) + c\mathbf{1}[\theta \notin a]$, *where* $\mu_L(a)$ *is the Lebesgue measure of the set* $a \subset \mathbb{R}$, *with* $d_J^*(y) = \{\theta : \phi_{\Sigma_J/n}(y - \theta) \geq 1/c\}$ *solving (7) (cf. Proposition 5.79 of Schervish (1995)). Also this loss function is invariant, since* $\ell(\theta, a) = \mu_L(a) + c\mathbf{1}[\theta \notin a] = \mu_L(q(-\theta, a)) + c\mathbf{1}[0 \notin$

$q(-\theta, a)]$, *where* $q(\theta, a) = \{t : t - \theta \in a\}$, *and* $d_J^*(y) = \{\theta : \phi_{\Sigma_J/n}(y - \theta) \geq 1/c\} = q(y, a_J^*)$, *where* $a_J^* = \{\theta : \phi_{\Sigma_J/n}(-\theta) \geq 1/c\}$.

2.3 Dependent Observations and Random Information

The discussion so far assumed that the observations x_i are independent draws from the a model with density f, and that the fitted model also assumes independent observations from the density g. But this restriction is not crucial. Let g_n and f_n be families of densities with respect to μ_n of the whole data vector $X_n = (x_1, \cdots, x_n)$, indexed by $\theta \in \Theta$—in the i.i.d. case, g_n and f_n are given by $g_n = g \times g \times \cdots \times g$ and $f_n = f \times f \times \cdots \times f$. Suppose for each θ_0, the Kullback-Leibler divergence between the true model f_n with parameter θ_0 and the fitted model g_n with parameter θ, $K_n(\theta) = \int f_n(\theta_0, X) \ln(f_n(\theta_0, X)/g_n(X, \theta)) d\mu_n(X)$, is minimized at θ_0, that is θ_0 is the pseudo-true value that maximizes the expected log likelihood $E \ln L_n(\theta) = E \ln g_n(X_n, \theta) = \int f_n(\theta_0, X) \ln g_n(X, \theta) d\mu_n(X)$. Let $L_i(\theta)$ and $S_i(\theta) = \partial L_i(\theta)/\partial\theta$ be the likelihood and scores of the first $i \leq n$ observations, define the differences $l_i(\theta) = L_i(\theta) - L_{i-1}(\theta)$, $s_i(\theta) = \partial l_i(\theta)/\partial\theta$ and $h_i(\theta) = \partial s_i(\theta)/\partial\theta'$. Under regularity conditions about the true model f_n, such as an assumption of $\{x_i\}$ to be stationary and ergodic, a (uniform) law of large numbers can be applied to $n^{-1} \sum_{i=1}^n h_i(\theta)$, justifying the quadratic approximations in (2) and (4). Furthermore, note that $\exp[l_i(\theta)] = g_i(X_i, \theta)/g_{i-1}(X_{i-1}, \theta)$ is the conditional density of x_i given X_{i-1} in the fitted model. In the correctly specified model with $f_n = g_n$, the scores $s_i(\theta_0)$ thus form a martingale difference sequence (m.d.s.) relative to the information $X_i = (x_1, \cdots, x_i)$, $E[s_i(\theta_0)|X_{i-1}] = 0$; cf. Chapter 6.2 of Hall and Heyde (1980). This suggests that in moderately misspecified models, $\{s_i(\theta_0)\}_{i=1}^n$ remains an m.d.s., or at least weakly dependent, so that an appropriate central limit theorem can be applied to $n^{-1/2} S_n(\theta_0) = n^{-1/2} \sum_{i=1}^n s_i(\theta_0)$. One would thus expect the arguments in Sections 2.1 and 2.2 to go through also for time series models.

A second generalization concerns the asymptotic variances Σ_S and Σ_M, which we assumed to be non-stochastic. Suppose instead of (3) and (4), the following convergences hold jointly

$$n^{1/2} \Sigma_S^{-1/2} (\hat{\theta} - \theta_0) \Rightarrow Z \sim \mathcal{N}(0, I_k), \; \hat{\Sigma}_S \xrightarrow{p} \Sigma_S \tag{11}$$

$$L_n(\hat{\theta} + n^{-1/2} u) - L_n(\hat{\theta}) \xrightarrow{p} -\tfrac{1}{2} u' \Sigma_M^{-1} u \tag{12}$$

where Σ_S and Σ_M are stochastic matrices that are positive definite with probability one, and Z is independent of (Σ_S, Σ_M). The log-likelihood that corresponds to the information

about θ embedded in $n^{1/2}\Sigma_S^{-1/2}(\hat{\theta}-\theta_0) \sim \mathcal{N}(0,I_k)$ equals $C-\frac{1}{2}n(\theta-\hat{\theta})'\Sigma_S^{-1}(\theta-\hat{\theta})$, while Bayesian inference in the misspecified model is based instead on the log likelihood $C-\frac{1}{2}n(\theta-\hat{\theta})'\Sigma_M^{-1}(\theta-\hat{\theta})$. Proceeding as in Section 2.2, the Gaussian model (6) is now replaced by the mixture model $\Sigma_S^{1/2}(Y-\theta) \sim \mathcal{N}(0,I_k/n)$, with large sample Bayesian inference in the misspecified model corresponding to best inference in the incorrect model $\Sigma_M^{1/2}(Y-\theta) \sim \mathcal{N}(0,I_k/n)$, and inference based on the sandwich likelihood $L_{Sn}(\theta)=C-\frac{1}{2}n(\theta-\hat{\theta})'\hat{\Sigma}_S^{-1}(\theta-\hat{\theta})$ corresponding to best inference in the correct model. The superiority of the latter then follows from the same reasoning as in the remainder of Section 2.2 by first conditioning on (Σ_S,Σ_M): frequentist risk in the mixture model is a weighted average of the frequentist risk given Σ_S, with weights corresponding to the probability distribution of Σ_S. Sandwich likelihood based inference is conditionally, and thus unconditionally best, and if $\Sigma_S \neq \Sigma_M$ with positive probability, Bayesian inference using the misspecified model is of higher risk for variance sensitive decisions.

Example 4 *Consider the linear regression model $y_i=z_i\theta+\varepsilon_i$ with $E[\varepsilon_i|z_i]=0$. Suppose $z_i=\xi z_i^*$, where $P(\xi=1)=1/2$ and $P(\xi=2)=1/2$, and $z_i^* \sim iid\mathcal{N}(0,1)$. Suppose the fitted model assumes standard normal disturbances independent of $\{z_i\}$, while the true model has homoskedastic disturbances of variance $\sigma^2 \neq 1$. The MLE is given by $\sqrt{n}(\hat{\theta}-\theta_0)=\hat{\Sigma}_M\left(n^{-1/2}\sum_{i=1}^n z_i\varepsilon_i\right)$, where $\hat{\Sigma}_M^{-1}=n^{-1}\sum_{i=1}^n z_i^2 \xrightarrow{p} \xi^2=\Sigma_M^{-1}$, and $n^{-1/2}\xi^{-1}\sum_{i=1}^n z_i\varepsilon_i = n^{-1/2}\sum_{i=1}^n z_i^*\varepsilon_i \Rightarrow \sigma Z \sim \mathcal{N}(0,\sigma^2)$, so that $\sqrt{n}\Sigma_S^{-1/2}(\hat{\theta}-\theta_0) \Rightarrow Z$ with $\Sigma_S=\sigma^2\xi^{-2}$. The log-likelihood in the fitted model satisfies (12), suggesting that the posterior for θ is large sample equivalent to the distribution $\theta \sim \mathcal{N}(\hat{\theta},\Sigma_M/n)$. The arguments of Section 2.2 now show that inference based on the sandwich posterior $\theta \sim \mathcal{N}(\hat{\theta},\hat{\Sigma}_S/n)$ with $\hat{\Sigma}_S=\hat{\Sigma}_M\left(n^{-1}\sum_{i=1}^n z_i^2e_i^2\right)\hat{\Sigma}_M \xrightarrow{p} \Sigma_S \neq \Sigma_M$ and $e=x-z_i\hat{\theta}$ is of lower risk conditional on ξ, and thus also unconditionally.*

2.4 Implementation Issues

The heuristic arguments so far suggest that if models are potentially misspecified, then lower risk decisions about pseudo-true values are obtained quite generally by basing inference on the artificial sandwich likelihood

$$L_{Sn}(\theta)=C-\tfrac{1}{2}n(\theta-\hat{\theta})'\hat{\Sigma}_S^{-1}(\theta-\hat{\theta}) \tag{13}$$

rather than the original likelihood. Implementation of this prescription requires the determination of $\hat{\theta}$ and $\hat{\Sigma}_S$.

In models with low dimensional θ, this is usually quite straightforward: The MLE can be obtained numerically, and $\hat{\Sigma}_S = \hat{\Sigma}_M \hat{V} \hat{\Sigma}_M$ with $\hat{\Sigma}_M^{-1} = n^{-1} \sum_{i=1}^n h_i(\hat{\theta})$ and $\hat{V} = n^{-1} \sum_{i=1}^n s_i(\hat{\theta}) s_i(\hat{\theta})'$ is often a precise estimator of Σ_S. In this approach, one could combine the sandwich likelihood (13) with the prior $p(\theta)$ to obtain sandwich corrected Bayesian inference. Alternatively, one might count on the likelihood to dominate the prior and exploit the output of a Bayesian posterior sampler of the misspecified model: Since the posterior distribution is approximately $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_M/n)$, one can directly use the mean and variance of the posterior as estimates of $\hat{\theta}$ and $\hat{\Sigma}_M/n$. The only additional piece required for the estimation of $\hat{\Sigma}_S$ then is $\hat{V} = n^{-1} \sum_{i=1}^n s_i(\hat{\theta}) s_i(\hat{\theta})'$.

In models with a high dimensional parameter, however, one might rightfully question the accuracy of quadratic approximations to the likelihood such as those in (2) and (4). The likelihood often is not very informative about all unknowns in the model, which can lead to numerical instability and implausible results from standard maximum likelihood estimation. In such models, it seems both necessary and sensible to impose some *a priori* knowledge about possible parameter values.

A potentially attractive option then is to maintain at least some of this *a priori* information, but to allow for other aspects of the model to be potentially misspecified. To fix ideas, suppose the parameters in the model are partitioned as (θ, γ), with primary interest in (a subset of) the $k \times 1$ vector θ. Let $L_n^c(\theta, \gamma)$ be the likelihood of the model given both parameters, let $p(\theta)$ be the prior over θ and $p^c(\gamma|\theta)$ be the prior on γ conditional on θ, so that the overall prior on (θ, γ) is given by $p(\theta, \gamma) = p(\theta) p^c(\gamma|\theta)$. The Bayes action then minimizes

$$\int \int \ell(\theta, a) \exp[L_n^c(\theta, \gamma)] p^c(\gamma|\theta) p(\theta) d\gamma d\theta = \int \ell(\theta, a) p(\theta) \left(\int \exp[L_n^c(\theta, \gamma)] p^c(\gamma|\theta) d\gamma \right) d\theta. \tag{14}$$

If the aim is to find a decision that minimize Bayes (=weighted average) risk, then it is irrelevant whether γ is thought of as a fixed but unknown parameter, or as stochastic with conditional density $p^c(\gamma|\theta)$: in the former case, the integration over p^c is part of the Bayes risk calculation, whereas in the latter, it is part of the frequentist risk. If γ is a "latent variable", such as an unobserved state, a stochastic view of γ is entirely natural also from a standard frequentist perspective; otherwise it would a "random parameter". Either way, in a correctly specified model, the decision that minimizes (14) for all realizations of X_n minimizes overall Bayes risk.

Taking the *a priori* information about γ embedded in $p^c(\gamma|\theta)$ seriously also for poten-

tially misspecified models leads to inference based on the sandwich correction (13) of the "integrated likelihood"

$$\exp[L_n(\theta)] = \int \exp[L_n^c(\theta,\gamma)]p^c(\gamma|\theta)d\gamma. \tag{15}$$

The function $L_n(\theta)$ in (15) is a proper log-likelihood under the stochastic interpretation for γ: it is the log-density of the model where γ is a random vector drawn from $p^c(\gamma|\theta)$, and X_n is then drawn from the density corresponding to $L_n^c(\theta,\gamma)$. Even if $L_n^c(\theta,\gamma)$ is the log-likelihood of i.i.d. data, data drawn from this model is almost never i.i.d., since the draw of γ is a common determinant for $x_1, \cdots, x_n$. What is more, the realization of the value of γ typically determines the information about θ, so that the random information generalization of the last subsection becomes pertinent.

Example 5 *Suppose the model and estimation methods are just as in Example 4, except for $z_i = \gamma z_i^*$, with some proper prior on γ that is independent of θ. The same discussion as in Example 4 then applies with $\xi = \gamma$.*

As before, $\hat{\theta}$ and $\hat{\Sigma}_M$ from model (15) can be numerically determined as the posterior mean and variance of a full sample Bayesian estimation with overall prior $p(\gamma,\theta)$ on the unknowns (θ,γ). It thus remains the issue of how to estimate the variance of the scores V. Just as in the discussion of the time series case in Section 2.3, with $S_i(\theta) = \partial L_i(\theta)/\partial\theta$, the differences

$$s_i(\theta_0) = S_i(\theta_0) - S_{i-1}(\theta_0)$$

form a m.d.s. relative to X_i in the correctly specified model. This suggests that also in a range of moderately misspecified models, $s_i(\theta_0)$ is a m.d.s. or at least uncorrelated, and a natural estimator for V is $\hat{V} = n^{-1}\sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})'$. The Bayesian computational machinery can now be put to use to numerically determine $\{S_i(\hat{\theta})\}$, and thus also $\{s_i(\hat{\theta})\}$: Straightforward calculus yields

$$\begin{aligned} S_i(\theta) &= \frac{\int \exp[L_i^c(\theta,\gamma)]p^c(\gamma|\theta)(\frac{\partial L_i^c(\theta,\gamma)}{\partial\theta} + \frac{\partial \ln p^c(\gamma|\theta)}{\partial\theta})d\gamma}{\int \exp[L_i^c(\theta,\gamma)]p^c(\gamma|\theta)d\gamma} \\ &= \int T_i^c(\theta,\gamma)d\Pi_i^c(\gamma|\theta) \end{aligned}$$

where $T_i^c(\theta,\gamma) = \partial L_i^c(\theta,\gamma)/\partial\theta + \partial \ln p^c(\gamma|\theta)/\partial\theta$, and $\Pi_i^c(\gamma|\theta)$ is the posterior distribution of γ in the model of X_i with log-likelihood $L_i^c(\theta,\gamma)$, conditional on θ. The vectors $S_i(\hat{\theta})$

can thus be determined by setting up a sampler for the posterior distribution of γ given the first i observations and $\theta = \hat{\theta}$, and by then computing the posterior weighted average of $T_i^c(\hat{\theta}, \gamma)$.

These considerations suggest the following approach to inference about pseudo-true parameter value θ in potentially misspecified Bayesian models.

Algorithm 1 *1. Compute the posterior mean $\hat{\theta}$ and variance $\hat{\Sigma}_M/n$ from a standard full sample Bayesian estimation with prior $p(\theta, \gamma)$ on the unknowns (θ, γ).*

2. For $i = 1, \cdots, n$, set up a posterior sampler for γ with prior $p^c(\gamma|\hat{\theta})$ conditional on $\theta = \hat{\theta}$ using the first i observations X_i only. Compute $S_i(\hat{\theta})$, the posterior mean of $T_i^c(\hat{\theta}, \gamma)$ under this sampler.

3. Base inference on the sandwich posterior $\mathcal{N}(\hat{\theta}, \hat{\Sigma}_S/n)$, where $\hat{\Sigma}_S = \hat{\Sigma}_M \hat{V} \hat{\Sigma}_M$, $\hat{V} = n^{-1} \sum_{i=1}^n s_i(\hat{\theta}) s_i(\hat{\theta})'$ and $s_i(\hat{\theta}) = S_i(\hat{\theta}) - S_{i-1}(\hat{\theta})$.

Comments:

1. The dimension of γ may depend on the sample size n, and subsets of γ may only affect a subset of observations. In this case only a subset of γ may be relevant for the estimation in Step 2 for some i, as a subset of γ does not enter $L_i^c(\theta, \gamma)$. Also, it might be that conditional on θ, there is no dependence between observations x_i and x_j either through the model $L_n^c(\theta, \gamma)$ or through the prior $p^c(\gamma|\theta)$. For instance, in a hierarchical panel model with n units, $\gamma = (\gamma_1, \cdots, \gamma_n)$ might contain individual specific parameters that are independent draws from a hierarchical prior parametrized by θ, $p^c(\gamma|\theta) = \prod_{i=1}^n p_i^c(\gamma_i|\theta)$, and $L_n^c(\theta, \gamma) = \sum_{i=1}^n l_i^c(\theta, \gamma_i)$. In this case, it suffices to re-estimate the model in Step 2 a single time on the whole data set conditional on $\theta = \hat{\theta}$: Observing x_i does not affect the posterior for γ_j, so that the differences $s_i(\hat{\theta})$ simply recover the posterior mean of the partial derivative $\partial l_i^c(\theta, \gamma_i)/\partial\theta + \partial \ln p^c(\gamma_i|\theta)/\partial\theta$, evaluated at $\theta = \hat{\theta}$.

2. More generally, though, Step 2 of the algorithm requires re-estimation of the model (conditional on $\theta = \hat{\theta}$) on an increasing subset of the observations. It is not important that the subsets increase by one observation at a time, nor does the order need to correspond to the recording of the observation. Any sequential revelation of the full data set X_n will do as long as (i) the scores $s_i(\theta_0)$ are uncorrelated also in the misspecified model and (ii) a law of large numbers provides a plausible approximation for $\hat{V}$. The former property depends on the type of model and form of misspecifications one is willing to entertain. For instance, for panel data or clustered data, misspecification might well lead to correlated

scores within a single unit, but treating the whole unit as one observation x_i preserves uncorrelatedness of the $s_i(\theta_0)$. Similar ideas can be applied in dynamically misspecified time series model by collecting adjacent observations in blocks.[1] In general, property (i) will be more easily satisfied when the number of revelation steps n is small.

Small n, however, will typically render the law of large number approximation (ii) less accurate. Especially when θ is a vector of reasonably high dimension k, one might think that n must be very large to obtain an accurate estimator of the $k \times k$ matrix $\hat{V}$. But if the primary object of interest is a (scalar) element $\iota'\theta$ of θ, where ι is the appropriate column of I_k, then its asymptotic variance is estimated by $\iota'\hat{\Sigma}_S\iota = n^{-1}\sum_{i=1}^n(\iota'\hat{\Sigma}_M s_i(\hat{\theta}))^2$, so that only a scalar Law of Large Numbers needs to provide a reasonably accurate approximation. (Of course, the quadratic approximation to the log-likelihood (4) underlying the posterior approximation $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_M/n)$ must also be reasonably good; but this does not depend on the number of revelation steps n used in Step 2 of the algorithm.) In addition, smaller n lessens the computational burden of Step 2: Not only because fewer samplers need to be run, but also because the numerical accuracy in the estimation of $S_i(\hat{\theta})$ needs to be very high if n is large, since Monte Carlo estimation error in the (then small) differences $s_i(\hat{\theta})$ will artificially inflate $\hat{V}$. This suggests that the algorithm might well be implemented most successfully for quite moderate n (say, $n = 25$). It is important, however, that that none of the $s_i(\theta_0)$ dominates the variability of $n^{-1/2}S_n(\theta_0) = n^{-1/2}\sum_{i=1}^n s_i(\theta_0)$. This rules out revelation schemes where one particular step reveals most of the information about θ in the data.

3. For some models and priors, it might be non-trivial to compute the partial derivative $\partial \ln p^c(\gamma|\theta)/\partial\theta$. But since the objective of the Algorithm is the estimation of the (conditional) variance V of

$$n^{-1/2}S_n(\theta_0) = n^{-1/2}\int \frac{\partial L_i^c(\theta,\gamma)}{\partial\theta}|_{\theta=\theta_0} d\Pi_n^c(\gamma|\theta_0) + n^{-1/2}\int \frac{\partial \ln p^c(\gamma|\theta)}{\partial\theta}|_{\theta=\theta_0} d\Pi_n^c(\gamma|\theta_0)$$

and the last term is usually $o_p(n^{-1/2})$ (unless the dimension of γ increases linearly in n), one can typically justify the additional approximation of ignoring the contribution of $\partial \ln p^c(\gamma|\theta)/\partial\theta$, and replace T_i^c by $\tilde{T}_i^c(\theta,\gamma) = \partial L_i^c(\theta,\gamma)/\partial\theta$.

Similarly, for some models, it might be non-trivial to set up a sampler that conditions on $\theta = \hat{\theta}$. But the overall posterior $\Pi_i(\theta,\gamma) \propto p(\theta,\gamma)\exp[L_i^c(\theta,\gamma)]$ of (θ,γ) based on the

[1] Alternatively, corrections of the Newey and West (1987)-type could be employed in the estimation of $\hat{V}$ from $\{s_i(\hat{\theta})\}$

observations X_i has a marginal for θ that concentrates on θ_0 for not too small i with high probability—after all, this is how $\hat{\theta}$ is determined in Step 1 with $i = n$. If the shape of the prior $p^c(\gamma|\theta)$ and likelihood $\exp[L_i^c(\theta, \gamma)]$ for γ is continuous as a function of θ at $\theta = \theta_0$, this suggests that if necessary, the sampler in Step 2 of the Algorithm can be replaced by a sampler that averages over $T_i^c(\gamma, \hat{\theta})$ with draws of γ from the posterior distribution $\Pi_i(\theta, \gamma)$, without conditioning on $\theta = \hat{\theta}$.

4. The choice of the partition of the model's unknowns in θ and γ is likely to be a trade-off between the extent of asymptotic robustness against many forms of misspecification and small sample approximation issues. If the likelihood is not very informative about a particular parameter, then the prior is not dominated, and quadratic approximations to the log-likelihood as in (2) and (4) are likely to be poor. It thus doesn't make sense to include such parameters in θ. On the other hand, by integrating out γ, inference about θ from model (15) will typically depend more crucially on the (partial) appropriateness of the model $L_n^c(\theta, \gamma)$ and prior $p^c(\gamma|\theta)$—as one extreme example, think of the case where $p^c(\gamma|\theta)$ equals zero at the pseudo-true value γ_0 of γ. At the time, note that if γ is strongly identified by the data and $p^c(\gamma_0|\hat{\theta}) > 0$, then the posterior distribution $\Pi_i^c(\gamma|\hat{\theta})$ of γ in Step 2 of the Algorithm concentrates on γ_0, at least for large enough values of i, so that $\int \partial L_i^c(\theta, \gamma)/\partial\theta|_{\theta=\hat{\theta}} d\Pi_i^c(\gamma|\hat{\theta}) \approx \partial L_i^c(\theta, \gamma_0)/\partial\theta|_{\theta=\hat{\theta}}$.

It is illustrative to consider the logic of the suggested algorithm in a case where the posteriors can be explicitly computed. To this end, consider an i.i.d. linear regression model with two scalar regressors

$$y_i = w_i\theta + z_i\gamma + \varepsilon_i$$

and $E[\varepsilon_i|w_i, z_i] = 0$ and $E[\varepsilon_i^2] = 1$. The fitted model assumes $\varepsilon_i \sim \mathcal{N}(0, 1)$ independent of (w_i, z_i) and independent standard normal priors on θ and γ. Here, $T_i^c(\theta, \gamma) = \sum_{j=1}^i (y_j - w_j\theta - z_j\gamma)w_j$, so that $S_i(\theta_0) = \sum_{j=1}^i (y_j - w_j\theta_0 - z_j\bar{\gamma}_i)w_j$, where $\bar{\gamma}_i = (1 + \sum_{j=1}^i z_j^2)^{-1} \sum_{j=1}^i z_j(y_j - \theta_0 w_j)$ is the posterior mean of γ in the fitted model based on the first i observations, conditional on $\theta = \theta_0$. Thus

$$s_i(\theta_0) = w_i\varepsilon_i + w_i z_i(\gamma - \bar{\gamma}_i) - (\bar{\gamma}_i - \bar{\gamma}_{i-1}) \sum_{j=1}^{i-1} w_j z_j \tag{16}$$

and tedious algebra but straightforward algebra shows $s_i(\theta_0)$ to be mean zero and uncorrelated if $E[\gamma] = 0$, $E[\gamma^2] = 1$ and $E[(z_j\varepsilon_j)^2|z_j, w_j] = z_j^2$, and a m.d.s. if additionally,

$E[\gamma|\{y_j, w_j, z_j\}_{j=1}^i] = \bar{\gamma}_i$ also in the misspecified model with $\theta = \theta_0$. Uncorrelatedness of $s_i(\theta_0)$ thus not only depends on the random parameter γ to have the same two moments as specified in its prior, but also that the conditional variance of ε_i given z_i does not depend on z_i. Thus, only heteroskedasticity mediated through w_i (but not z_i) is generically compatible with uncorrelated $s_i(\theta_0)$. At the same time, since $n(\bar{\gamma}_n - \bar{\gamma}_{n-1}) = z_n\varepsilon_n/E[z_i^2] + o_p(1)$, it follows from (16) that $s_i(\theta_0) = (w_i - z_i E[w_i z_i]/E[z_i^2])\varepsilon_i + r_i$ with $r_{i_n} \xrightarrow{p} 0$ conditional on γ for any sequence $i_n \to \infty$. The same holds for $s_i(\hat{\theta})$, so in large samples and irrespective of the true value of γ, the algorithm yields an estimator $\hat{\Sigma}_S$ that is equivalent to the usual heteroskedasticity robust variance estimator for θ, allowing for heteroskedasticity mediated through both w_i and z_i.

3 Large Sample Risk Comparisons

This Section formalizes the heuristics of Section 2. The first subsection introduces the main assumptions. The second subsection formally states the large sample superiority of basing inference about pseudo-true parameter values on the sandwich likelihood in misspecified Bayesian models, followed by a discussion. Lastly, we present more primitive conditions that are shown to imply the main condition.

3.1 Assumptions

The observations in a sample of size n are vectors $x_i \in \mathbb{R}^r$, $i = 1, \cdots, n$, with the whole data denoted $X_n = (x_1, \cdots, x_n)$, and the model with log-likelihood function $L_n : \Theta \times \mathbb{R}^{r \times n} \mapsto \mathbb{R}$ 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 parametrized such that θ_0 is pseudo-true relative to the assumed model, that is, $\int L_n(\theta_0, X) dP_{n,\theta_0}(X) = \sup_\theta \int L_n(\theta, X) dP_{n,\theta_0}(X)$ for all $\theta \in \Theta$. The prior on $\theta \in \Theta$ is described by the Lebesgue density p, and the data-dependent posterior computed from a potentially misspecified parametric model with parameter θ is denoted by Π_n. Let $\hat{\theta}$ be an estimator of θ (such as the MLE), and let $d_{TV}(P_1, P_2)$ be the total variation distance between the 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) $\sqrt{n}\Sigma_S(\theta_0)^{-1/2}(\hat{\theta}-\theta_0) \Rightarrow Z$ *with* $Z \sim \mathcal{N}(0, I_k)$ *and there exists an estimator* $\hat{\Sigma}_S \xrightarrow{p} \Sigma_S(\theta_0)$*, where* $\Sigma_S(\theta_0)$ *is independent of* Z *and positive definite almost surely.*

(ii) $d_{TV}(\Pi_n, \mathcal{N}(\hat{\theta}, \Sigma_M(\theta_0)/n)) \xrightarrow{p} 0$*, where* $\Sigma_M(\theta_0)$ *is independent of* Z *and positive definite almost surely.*

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 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 are provided by Bunke and Milhaud (1998) and Kleijn and van der Vaart (2008) in models with i.i.d. observations, and the general results of Chen (1985) can be used to establish the convergence also in the non-i.i.d. case. Section 3.3 below provides more primitive assumptions that lead to Condition 1 also for stochastic $\Sigma_M(\theta_0)$ 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* $a \in \mathcal{A}$*,* $\theta \in \mathbb{R}^k$*.*

Condition 2 restricts the loss to be non-negative and bounded. Bounded loss ensures that small probability events only have a small effect on overall risk, which allows precise statements in combination of 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.

Example 6 *Replacing the squared loss in Example 1 by truncated squared loss* $\ell(\theta, a) = \min((\theta-a)^2, \bar{\ell})$ *for* $\bar{\ell} > 0$ *still yields the same Bayes action* $d_J^*(y) = (\mu_p \Sigma_J/n + \sigma_p^2 y)/(\Sigma_J/n +$

σ_p^2) *under a normal posterior by Anderson's Lemma. For a bounded linex loss* $\ell(\theta, a) = \min(\exp[b(\theta - a)] - b(\theta - a) - 1, \bar{\ell})$ *and bounded interval estimation loss* $\ell(\theta, a) = \min(a_u - a_l + c(\mathbf{1}[\theta_1 < a_l](a_l - \theta_{(1)}) + \mathbf{1}[\theta > a_u](\theta_{(1)} - a_u), \bar{\ell})$ *in Example 2, a calculation shows that the optimal decisions (7) approach the solutions to the untruncated problem as* $\bar{\ell} \to \infty$.

In the general setting with data $X_n \in \mathbb{R}^{r \times n}$, decisions d_n are measurable mappings 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 are given by

$$\begin{aligned} r_n(\theta, d_n) &= \int \ell(\theta, d_n(X)) dP_{n,\theta}(X) \\ R_n(p, d_n) &= \int r_n(\theta, d_n) p(\theta) d\theta \end{aligned}$$

respectively.

The motivation for allowing sample size dependent loss functions is not necessarily that more data leads 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 n, asymptotically nontrivial decisions 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 square 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 estimators with smaller asymptotic variance is relatively smaller for large n, and a corresponding formal result is obtained by choosing $\ell_n(\theta, a) = \min(n(\theta - a)^2, \bar{\ell})$.

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

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

for almost all X_n. As discussed in the heuristic section, it makes sense to compare the performance of d_{Mn} with the decision rules that are computed from the "sandwich" posterior

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

In particular, suppose d_{Sn} satisfies

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

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

A first and maybe most attractive result is obtained under the following condition about the loss function.

Condition 3 *(i) ℓ_n is asymptotically locally invariant at θ_0, that is*

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

for some measurable function $\ell_n^i : \mathbb{R}^k \mapsto \mathbb{R}$ with $\ell_n^i(\theta, a) = \ell_n^i(0, q(-\theta, a))$ for all $a \in \mathcal{A}$, $\theta \in \mathbb{R}^k$ and some function $q : \mathbb{R}^k \times \mathcal{A} \mapsto \mathcal{A}$ satisfying $q(\theta_1, q(\theta_2, a)) = q(\theta_1 + \theta_2, a)$ for all $\theta_1, \theta_2 \in \Theta$.

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 $\ell_n^i(u_n/\sqrt{n}, a_n) - \ell_n^i(u/\sqrt{n}, a_n^*(\Sigma_J(\theta_0))) \to 0$.

Condition 3 (i) assumes that at least in the $\sqrt{n}$-neighborhood of θ_0, the loss functions ℓ_n are well approximated by invariant loss functions ℓ_n^i for large enough n. Parts (ii) and (iii) assume a two-fold continuity of ℓ_n^i: For $J = S, M$, if a sequence of actions a_n comes close to minimizing risk relative to $\mathcal{N}(0, \Sigma_J(\theta_0)/n)$, then (a) it yields similar losses as the optimal actions $a_n^*(\Sigma_J(\theta_0))$, $\ell_n^i(u/\sqrt{n}, a_n) - \ell_n^i(u/\sqrt{n}, a_n^*(\Sigma_J(\theta_0))) \to 0$ for almost all u, and (b) losses incurred along the sequence $u_n \to u$ are close to those obtained at u, $\ell_n^i(u_n/\sqrt{n}, a_n) - \ell_n^i(u/\sqrt{n}, a_n) \to 0$. These are non-trivial restrictions on the loss functions. But to reduce the asymptotic problem decision problem to the normal location problem of Section 2.2 under Condition 1, one must ensure that the small differences between the sampling distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ and $\mathcal{N}(0, \Sigma_S(\theta_0))$, and of Π_n and $\mathcal{N}(\hat{\theta}, \Sigma_M(\theta_0)/n)$, cannot lead to substantially different risks.

Example 7 *Consider the interval estimation problem with $\ell_n(\theta, a) = \min(\sqrt{n}(a_u - a_l + c\mathbf{1}[\theta_1 < a_l](a_l - \theta_{(1)}) + c\mathbf{1}[\theta > a_u](\theta_{(1)} - a_u)), \bar{\ell})$ and large $\bar{\ell}$, where the scaling by*

$\sqrt{n}$ *prevents that all reasonable decisions have zero asymptotic risk. Assume* $\Sigma_J(\theta_0)$ *is almost surely constant, so that* $a^*_{Jn} = a^*_n(\Sigma_J(\theta_0))$ *is not random. Then* $\sqrt{n}a^*_{Jn} = (-z_c\kappa_{\bar{\ell}}\sqrt{\Sigma_{J(1,1)}}, z_c\kappa_{\bar{\ell}}\sqrt{\Sigma_{J(1,1)}})$ *with* $\kappa_{\bar{\ell}} < 1$ *a correction factor for the fact that loss is bounded, and any sequence* a_n *that satisfies the premise of part (iii) of Condition 3 must satisfy* $\sqrt{n}a_n - \sqrt{n}a^*_{Jn} \to 0$. *Thus* $\ell^i_n(u_n/\sqrt{n}, a_n) - \ell^i_n(u/\sqrt{n}, a^*_{Jn}) \to 0$ *for all* $u \in \mathbb{R}^k$, *so Condition 3 holds.*

In contrast, consider the set estimation problem with $\mathcal{A} = \{$*all Borel subsets of* $\mathbb{R}^k\}$ *and* $\ell_n(\theta, a) = \min(\sqrt{n}(\mu_L(a) + c\mathbf{1}[\theta \notin a]), \bar{\ell})$ *with* $\bar{\ell}$ *large. It is quite preposterous, but nevertheless compatible with Condition 1 (ii) that the posterior* Π_n *has a density that essentially looks like* $\phi_{\Sigma_M/n}(\theta - \hat{\theta})$, *but with an additional extremely thin (say, of base volume* n^{-2}*) and very high (say, of height* n*) peak around* θ_0, *almost surely. If that was the case, then* d_{Mn} *would, in addition to the HPD region computed from* $\phi_{\Sigma_M/n}(\theta - \hat{\theta})$, *include a small additional set of measure* n^{-2} *that always contains the true value* θ_0. *The presence of that additional peak induces a substantially different risk. It is thus not possible to determine the asymptotic risk of* d_{Mn} *under Condition 1 in this decision problem, and correspondingly,* $\ell_n(\theta, a) = \min(\sqrt{n}(\mu_L(a) + c\mathbf{1}[\theta \notin a]), \bar{\ell})$ *does not satisfy Condition 3.*[2] *In the same decision problem with the action space restricted to* $\mathcal{A} = \{$*all convex subsets of* $\mathbb{R}^k\}$, *however, the only actions that satisfy the premise of part (iii) in Condition 3 converge to* a^*_{Jn} *in the Hausdorff distance, and* $\ell^i_n(u_n/\sqrt{n}, a_n) - \ell^i_n(u/\sqrt{n}, a^*_{Jn}) \to 0$ *holds for all* u *that are not on the boundary of* a^*_{Jn}.

In absence of a (local) invariance property of ℓ_n, it is necessary to consider the properties of the stochastic 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 4 *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* $\Sigma_J(\cdot)$, $\int \ell_n(\theta, d^*_n(y, \Sigma_J(y)))\phi_{\Sigma_J(y)/n}(\theta - y)d\theta = \inf_{a\in\mathcal{A}} \int \ell_n(\theta, a)\phi_{\Sigma_J(y)/n}(\theta - y)d\theta$ *for* η*-almost all* $y \in \Theta$;

[2]Suppose $k = 1$, pick K large enough such that $n^{-1/2}(K-1) \notin a^*_{Jn} = \{\theta : \phi_{\Sigma_J(\theta_0)/n}(-\theta) \geq 1/c\}$, and define the intervals $I_n = [\sin(\ln n) - n^{-1/2}, \sin(\ln n) + n^{-1/2}]$. Note that I_n cover all numbers in the interval $[-1, 1]$ infinitely often, yet $\mu_L(I_n) = 2n^{-1/2} \to 0$. Thus, $a_n = a^*_{Jn} \cup \{\theta : n^{-1/2}\theta - K \in I_n\}$ violates condition (iii) for all $u \in [K-1, K+1]$.

(iii) for η-almost all θ_0, P-almost all $\Sigma_J(\cdot)$ and Lebesgue almost all $u \in \mathbb{R}^k$: $\int \ell_n(\theta, a_n)\phi_{\Sigma_J(y_n)/n}(\theta - y_n)d\theta - \int \ell_n(\theta, d_n^*(y_n, \Sigma_J(y_n)))\phi_{\Sigma_J(y_n)/n}(\theta - y_n)d\theta \to 0$ and $\sqrt{n}(y_n - \theta_0) \to u$ for some sequences $a_n \in \mathcal{A}$ and $y_n \in \mathbb{R}^k$ 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 (7) of Section 2. 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 (7) would naturally minimize $\int \ell_n(\theta, a)\phi_{\Sigma_J(\theta)/n}(\theta - y)d\theta$, whereas the assumption in part (ii) assumes d_n^* to minimize the more straightforward Gaussian problem with known covariance matrix $\Sigma_J(y)/n$. The proof of Theorem 2 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 with known covariance matrix $\hat{\Sigma}_S$.

Part (iii) of Condition 4 is similar to Condition 3 (iii) discussed above: If a sequence of a_n comes close to minimizing the same risk as $d_n^*(y_n, \Sigma_J(y_n))$ for some y_n satisfying $\sqrt{n}(y_n - \theta_0) \to u$, 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.

3.2 Main Result and Discussion

The proof of the following Theorem is in the appendix.

Theorem 2 *(i) Under Conditions 1, 2 and 3,*

$$\begin{aligned} r_n(\theta_0, d_{Mn}) - E[\int \ell_n^i(\theta, a_n^*(\Sigma_M(\theta_0))\phi_{\Sigma_S(\theta_0)/n}(\theta)d\theta] &\to 0 \\ r_n(\theta_0, d_{Sn}) - E[\int \ell_n^i(\theta, a_n^*(\Sigma_S(\theta_0))\phi_{\Sigma_S(\theta_0)/n}(\theta)d\theta] &\to 0. \end{aligned}$$

(ii) Under Conditions 2 and 4,

$$\begin{aligned} R_n(\eta, d_{Mn}) - E[\int\int \ell_n(\theta, d_n^*(y, \Sigma_M(y)))\phi_{\Sigma_S(y)/n}(\theta - y)d\theta\ \eta(y)dy] &\to 0 \\ R_n(\eta, d_{Sn}) - E[\int\int \ell_n(\theta, d_n^*(y, \Sigma_S(y)))\phi_{\Sigma_S(y)/n}(\theta - y)d\theta\ \eta(y)dy] &\to 0. \end{aligned}$$

1. The results in the two parts of Theorem 2 mirror the heuristics of Sections 2.2.1-2.2.2 above: For non-stochastic 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 likelihood (18) based decision d_{Sn} behave just like in the Gaussian location problem discussed there. In particular, this immediately implies that the decisions d_{Sn} are at least as good as d_{Mn} in large samples—formally, the two parts of Theorem 2 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 2.2, 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 and often better as d_{Mn}, the overall quality of the decisions d_{Sn} depends both on the relationship between the misspecified model and the true model, and how one defines "overall quality". For simplicity, we assume the asymptotic variances to be constant in the following discussion.

First, suppose the data generating process is embedded in a correct parametric model with true parameter $\theta_0 \in \Theta$. 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 2.1, under a smooth prior, the correct posterior Π_{Cn} converges to the distribution $\mathcal{N}(\hat{\theta}_C, \Sigma_C(\theta_0)/n)$, and $\hat{\theta}_C$ has the 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 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 $\Sigma_S(\theta_0) = \Sigma_C(\theta_0)$, 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 of a "corrected" posterior with 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 covariance matrix $\hat{\Sigma}_S$ in the definition of d_{Sn} yields the decision with the smallest large sample risk, and d_{Sn} might be considered optimal in this sense. Formally, in the context of the approximately invariant loss of Condition 3, consider the class of decisions $d_{Qn} : \mathbb{R}^k \mapsto \mathcal{A}$, indexed by $Q \in \mathcal{P}^k$, that satisfy

$$\int \ell_n(\theta, d_{Qn}(\hat{\theta}))\phi_{Q/n}(\theta - \hat{\theta})d\theta = \inf_{a\in\mathcal{A}} \int \ell_n(\theta, a)\phi_{Q/n}(\theta - \hat{\theta})d\theta.$$

If Condition 3 is strengthened to hold also for Q in place of $\Sigma_J(\theta_0)$, then proceeding as in the proof of Theorem 2 yields $r_n(\theta_0, d_{Qn}) - \int \ell_n^i(\theta, a_n^*(Q)\phi_{\Sigma_S(\theta_0)/n}(\theta - \theta_0)d\theta \to 0$, 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 variance adjustment to the posterior of a potentially misspecified model employs the sandwich matrix. This is true whether or not the adjusted 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) argue on page 402 "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 *adjustment* to the posterior computed from the misspecified model employs the sandwich matrix.

Related to this point is the approach of Kwan (1999). Kwan considers "limited information" posteriors that arise through conditioning on a statistic, such as an estimator $\tilde{\theta}$, rather than on the whole data vector X_n. He seeks to establish conditions under which convergence in distribution of $\sqrt{n}(\tilde{\theta} - \theta_0)$ implies a corresponding convergence in distribution of the limited information posterior. An application of his Theorem 1 would imply that the limited information posterior for the MLE $\hat{\theta}$ converges in distribution to $\mathcal{N}(\hat{\theta}, \Sigma_S(\theta_0)/n)$, as long as $\hat{\theta}$ is a regular estimator and the prior density is continuous and positive at the true parameter value. The decision d_{Sn} could then possibly be characterized as the best decision given the limited information contained in $\hat{\theta}$. Kwan's results are incorrect in the stated generality, however.[3] What is more, as demonstrated by Example 7, even the convergence in total variation of the posterior (which is stronger than convergence in distribution) does not necessarily imply that decisions computed from the limiting normal posterior have similar risk as decisions computed from the exact posterior, even asymptotically. Finally, given that the Bayes action relative to the approximate posterior $\mathcal{N}(\hat{\theta}, \Sigma_S(\theta_0)/n)$ typically depends on the unknown $\Sigma_S(\theta_0)$, it would be necessary to establish the limited information posterior of the pair $(\hat{\theta}, \hat{\Sigma}_S)$, which leads to further complications.

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 Normal linear regression model, 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

[3]Let $\hat{\theta}_n$ be generated as follows: the first n digits in the decimal expansion of $\hat{\theta}_n$ are equal to the corresponding digits of $\theta + Z/\sqrt{n}$, where $\theta \in \Theta = (0, 1)$ and $Z \sim \mathcal{N}(0, 1)$, and the remaining digits of $\hat{\theta}_n$ contain the complete decimal expansion of θ. With a uniform prior, the conditional distribution of θ given $\hat{\theta}_n$ is then degenerate at θ for all $n \geq 1$, yet the conditions of Theorem 1 in Kwan (1999) are satisfied.

error variance does not depend on 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, it is not possible to rank the relative performance of inference based on the original, misspecified model and inference based on sandwich posterior (18).

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 (2007) provide 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 as outlined in Section 2 above). 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 problem of parameter interval estimation under the loss described in Example 7, the practical implication of Theorem 2 part (i) is to report the standard frequentist confidence interval. The large sample equivalence of Bayesian and frequentist interval estimation in correctly specified models thus extends to a large sample equivalence of risk minimizing and frequentist interval estimation in moderately misspecified models that satisfy Condition 1.

This equivalence, however, only arises because the shape of the likelihood is identical to the density of the sampling distribution in the limiting Gaussian location problem. In less standard problems, the risk of interval estimators is still improved by taking into account the actual sampling distribution of the MLE, but without obtaining a large sample equivalence to the frequentist confidence interval. For concreteness, consider Sims and Uhlig's (1991) example of inference about the coefficient in a Gaussian autoregressive process of

order one with parameter values close to unity

$$x_i = \rho x_{i-1} + \varepsilon_i, \quad \varepsilon_i \sim iid\mathcal{N}(0, \sigma^2),\ x_0 = 0.$$

Suppose this model is potentially misspecified, with the true model of the form $x_i = \rho x_{i-1} + u_i$, and u_i is mean-zero stationary with variance σ^2 and long-run variance equal to $\bar{\sigma}^2$. Under local-to-unity asymptotics with $\rho_n = 1 - c/n$ and true value of c equal to c_0, the log-likelihood in the i.i.d. model has the shape

$$\begin{aligned} L_n(\rho_n) - L_n(1) &= (1-\rho_n)\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{i-1}\Delta x_i - \tfrac{1}{2}(1-\rho_n)^2\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{i-1}^2 \\ &\Rightarrow\ c\frac{\bar{\sigma}^2}{\sigma^2}\int J_{c_0}(s)dJ_{c_0}(s) + \tfrac{1}{2}c(\frac{\bar{\sigma}^2}{\sigma^2} - 1) - \tfrac{1}{2}c^2\frac{\bar{\sigma}^2}{\sigma^2}\int J_{c_0}(s)^2 ds \end{aligned} \tag{20}$$

where $J_c(s) = \int_0^s e^{-c(s-r)}dW(r)$ and W is a standard Wiener process, and the MLE satisfies (cf. Phillips (1987))

$$n(\hat{\rho}_n - 1) \Rightarrow \frac{\bar{\sigma}^2 \int J_{c_0}(s)dJ_{c_0}(s) + \frac{1}{2}(\bar{\sigma}^2 - \sigma^2)}{\bar{\sigma}^2 \int J_{c_0}(s)^2 ds}. \tag{21}$$

In the correctly specified model, $\bar{\sigma}^2 = \sigma^2$, and the term $\frac{1}{2}c(\bar{\sigma}^2 - \sigma^2)$ is not present in (20) or (21). Now suppose there are consistent estimators $\widehat{\bar{\sigma}}^2$ and $\hat{\sigma}^2$ of $\bar{\sigma}^2$ and σ^2. Since Bayesian inference based on the correct likelihood is Bayes risk minimizing, and the adjusted likelihood L_{Sn}

$$\begin{aligned} L_{Sn}(\rho_n) - L_n(1) &= (1-\rho_n)\left[\frac{1}{\widehat{\bar{\sigma}}^2}\sum_{i=1}^{n} x_{i-1}\Delta x_i - \tfrac{1}{2}n(1 - \frac{\hat{\sigma}^2}{\widehat{\bar{\sigma}}^2})\right] - \tfrac{1}{2}(1-\rho_n)^2\frac{1}{\widehat{\bar{\sigma}}^2}\sum_{i=1}^{n} x_{i-1}^2 \\ &\Rightarrow\ c\int J_{c_0}(s)dJ_{c_0}(s) - \tfrac{1}{2}c^2\int J_{c_0}(s)^2 ds \end{aligned}$$

has the same limiting behavior as the likelihood in a correctly specified model with $\sigma^2 = \bar{\sigma}^2$, one obtains asymptotically better inference using the adjusted likelihood L_{Sn} over the original likelihood L_n. Because L_{Sn} is quadratic in $(\rho_n - 1)$, under an approximately flat prior on c, the Bayes risk minimizing interval estimator for a loss function that rationalizes the 95% posterior probability interval based on L_{Sn} is given by

$$[\hat{\rho}_{Sn} - 1.96\left(\frac{1}{\widehat{\bar{\sigma}}^2}\sum_{i=1}^{n} x_{i-1}^2\right)^{-1/2}, \hat{\rho}_{Sn} + 1.96\left(\frac{1}{\widehat{\bar{\sigma}}^2}\sum_{i=1}^{n} x_{i-1}^2\right)^{-1/2}]$$

where $\hat{\rho}_{Sn} = \hat{\rho}_n - \frac{1}{2}n(\widehat{\bar{\sigma}}^2 - \hat{\sigma}^2)/\sum_{i=1}^{n} x_{i-1}^2$. But $\hat{\rho}_{Sn}$ does not have a Gaussian limiting distribution, so this interval is *not* a 95% confidence interval, even asymptotically.

3.3 Justification of Condition 1

The following Theorem provides more primitive assumptions that are sufficient for Condition 1, using notation established in Section 2. The result also holds for double array processes. The proof is in the appendix.

Theorem 3 *If under* P_{n,θ_0}

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

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

(iii) $\sup_{i\leq n} n^{-1/2}||s_i(\theta_0)|| \xrightarrow{p} 0$, $n^{-1}\sum_{i=1}^n s_i(\theta_0)s_i(\theta_0)' \xrightarrow{p} V(\theta_0)$, *where* $V(\theta_0) \in \mathcal{P}^k$ *almost surely, and* $n^{-1/2}\sum_{i=1}^n s_i(\theta_0) \Rightarrow V(\theta_0)^{1/2}Z$ *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* $P_{n,\theta_0}(\sup_{||\theta-\theta_0||\geq\epsilon} n^{-1}(L_n(\theta) - L_n(\theta_0)) < -K(\epsilon)) \to 1$*;*

(v) $n^{-1}\sum_{i=1}^n ||h_i(\theta_0)|| = O_p(1)$, *for any null sequence* k_n, $\sup_{||\theta-\theta_0||<k_n} n^{-1}\sum_{i=1}^n ||h_i(\theta) - h_i(\theta_0)|| \xrightarrow{p} 0$, *and* $n^{-1}\sum_{i=1}^n h_i(\theta_0) \xrightarrow{p} -\Sigma_M^{-1}(\theta_0)$, *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 $\hat{\Sigma}_S = \hat{\Sigma}_M\hat{V}\hat{\Sigma}_M$,$\hat{V} = n^{-1}\sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})'$ *and either (a)* $\hat{\theta}$ *equal to the MLE and* $\hat{\Sigma}_M^{-1} = -n^{-1}\sum_{i=1}^n h_i(\hat{\theta})$ *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, $s_i(\theta_0)$ forms a martingale difference sequence, then the last assumption in part (iii) holds if $\max_{i\leq n} E[s_i(\theta_0)'s_i(\theta_0)] = O(1)$ by Theorem 3.2 of Hall and Heyde (1980) and the so-called Cramer-Wold device. Alternatively, in the context of Algorithm 1 of Section 2.4, one might be able to argue that conditional on an appropriate subset $\gamma_{(1)}$ of γ, $\{s_i(\theta_0)\}$ can be well approximated by a zero mean, weakly dependent and uncorrelated series with (average) long-run variance that only depends on $\gamma_{(1)}$. The convergences in part (iii) can then be established by invoking an appropriate law of large numbers and central limit theorem for weakly dependent series conditional on $\gamma_{(1)}$. Assumption (iv) is the identification condition employed by Schervish (1995), page 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. Assumption (v) are fairly standard regularity conditions about the Hessians which can be established using the general results in Andrews (1987), possibly by again first conditioning

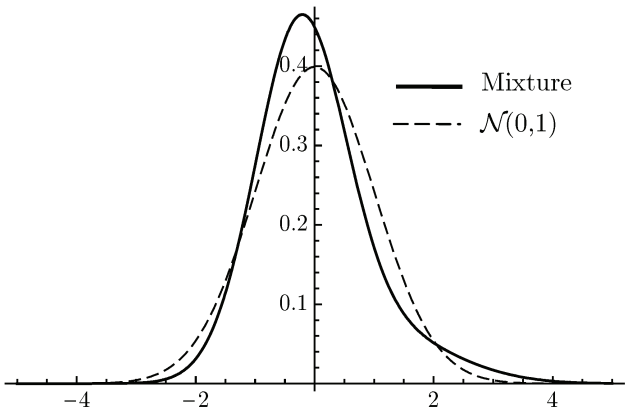


Figure 1: Asymmetric Mixture-of-Two-Normals Density

on an appropriate subset of γ. If the posterior is known to have at least two moments, then conclusion (b) of Theorem 3 also holds for $\hat{\theta}$ equal to the posterior mean and $\hat{\Sigma}_M$ equal to n times the posterior variance.

4 Small Sample Results

As a numerical illustration, consider a linear regression model with $\theta = (\alpha, \beta)'$ and one non-constant regressor z_i,

$$y_i = \alpha + z_i\beta + \varepsilon_i, \quad (y_i, z_i) \sim i.i.d., \;\; i = 1, \cdots, n. \tag{22}$$

We only consider data generating processes with $E[\varepsilon_i|z_i] = 0$, and assume throughout that the parameter of interest is given by $\beta \in \mathbb{R}$, the population regression coefficient. 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 z_i by one unit. Also, by construction, $\alpha + z_i\beta$ is the best predictor for $y_i|z_i$ under squared loss. Alternatively, a focus on β might be justified because economic theory implies $E[\varepsilon_i|z_i] = 0$. 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 $y_i|z_i$ under absolute value loss is the median of $y_i|z_i$, which does not coincide with the population regression function $\alpha + z_i\beta$ in general.

We consider six particular data generating processes (DGPs) satisfying (22). In all of them, $z_i \sim \mathcal{N}(0, 1)$. The first DGP is the normal linear model (DNLR) with $\varepsilon_i|z_i \sim \mathcal{N}(0, 1)$.

The second model has an error term that is a mixture (DMIX) of two normals where $\varepsilon_i|z_i, s \sim \mathcal{N}(\mu_s, \sigma_s^2)$, $P(s=1)=0.8$, $P(s=2)=0.2$, $\mu_1=-0.25$, $\sigma_1=0.75$, $\mu_2=1$ 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 the mixture model, but introduces a conditional asymmetry (DCAS) as a function of the sign of z_i: $\varepsilon_i|z_i, s \sim \mathcal{N}((1-2\cdot\mathbf{1}[z_i<0])\mu_s, \sigma_s^2)$. The final three DGPs are heteroskedastic versions of these homoskedastic DGPs, where $\varepsilon_i|z_i, s = a(0.5+|z_i|)\varepsilon_i^*$, ε_i^* is the disturbance of the homoskedastic DGP, and $a=0.454\ldots$ is the constant that ensures $E[(z_i\varepsilon_i)^2]=1$.

Inference is based on one of the following three methods. First, Bayesian inference with the normal linear regression model (INLR) where $\varepsilon_i|z_i \sim \mathcal{N}(0, h^{-1})$, with priors $(\alpha,\beta)' \sim \mathcal{N}(0, 100I_2)$ and $3h \sim \chi_3^2$. Second, Bayesian inference with a normal mixture linear regression model (IMIX), where $\varepsilon_i|z_i, s \sim \mathcal{N}(\mu_s, (hh_s)^{-1})$, $P(s=j)=\pi_j$, $j=1,2,3$, with priors $(\alpha,\beta)' \sim \mathcal{N}(0,100I_2)$, $3h \sim \chi_3^2$, $3h_j \sim iid\chi_3^2$, (π_1,π_2,π_3) ~Dirichlet$(3,3,3)$ and $\mu_j|h$ ~i.i.d.$\mathcal{N}(0, 2.5h^{-1})$. Third, inference based on the artificial sandwich posterior $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_S/n)$ from the normal linear regression model as suggested by Theorem 2 part (i) (ISAND), where $\hat{\theta}$ is the ordinary least squares coefficient, $\hat{\Sigma}_S = \hat{\Sigma}_M\hat{V}\hat{\Sigma}_M$, $\hat{\Sigma}_M = \hat{h}_n^{-1}\left(n^{-1}\sum_{i=1}^n Z_iZ_i'\right)^{-1}$, $\hat{V} = n^{-1}\hat{h}_n^2\sum_{i=1}^n Z_iZ_i'e_i^2$, $\hat{h}_n^{-1} = n^{-1}\sum_{i=1}^n e_i^2$, $Z_i=(1,z_i)$ and $e_i = y_i - Z_i'\hat{\theta}$.

Table 1 provides risks of Bayesian inference based on INLR and IMIX relative to ISAND at $\alpha=\beta=0$ for the scaled and bounded linex loss

$$\ell_n(\theta,a) = \min(\exp[2\sqrt{n}(\beta-a)] - 2\sqrt{n}(\beta-a) - 1, 80) \tag{23}$$

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

$$\ell_n(\theta,a) = \min(\sqrt{n}(a_u-a_l+40\cdot\mathbf{1}[\beta<a_l](a_l-\beta)+40\cdot\mathbf{1}[\beta>a_u](\beta-a_u)), 200) \tag{24}$$

with $a=(a_l,a_u)\in\mathbb{R}^2$, $a_u\geq a_l$, respectively. The bounds are approximately 40 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, INLR is slightly better than ISAND under homoskedasticity, with a somewhat more pronounced difference in the other direction under heteroskedasticity. This is not surprising, as INLR 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

Table 1: Risk of Decisions about Linear Regression Coefficient

	homoskedasticity			heteroskedasticity		
	DNLR	DMIX	DCAS	DNLR	DMIX	DCAS
Linex Loss, $n = 50$						
INLR	0.95	1.00	0.89	1.15	1.22	1.02
IMIX	0.98	0.89	0.93	1.03	0.85	1.20
Linex Loss, $n = 200$						
INLR	0.98	0.99	0.93	1.26	1.33	1.17
IMIX	1.01	0.83	1.47	1.17	0.87	3.78
Linex Loss, $n = 800$						
INLR	1.01	1.00	0.97	1.32	1.34	1.25
IMIX	1.01	0.84	5.16	1.23	0.91	16.6
Interval Estimation Loss, $n = 50$						
INLR	0.97	0.99	0.97	1.18	1.19	1.18
IMIX	0.98	0.92	1.00	1.06	0.95	1.13
Interval Estimation Loss, $n = 200$						
INLR	0.98	0.99	0.99	1.29	1.33	1.33
IMIX	1.00	0.90	1.23	1.17	1.04	2.45
Interval Estimation Loss, $n = 800$						
INLR	1.00	1.00	0.99	1.35	1.35	1.35
IMIX	1.00	0.90	2.66	1.22	1.08	8.07

Notes: Data generating processes are in columns, modes of inference in rows. Entries are the risk under linex loss (23) and interval estimation loss (24) of Bayesian inference using the row model divided by the risk of Bayesian inference using the artificial sandwich posterior $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_n/n)$. Risks are estimated from 5,000 draws from each DGP. Bayesian inference in INLR and IMIX is implemented by a Gibbs sampler with 7,000 draws, with 2,000 discarded as burn-in, and the Bayes actions are numerically determined from the resulting estimate of the posterior distribution.

better than ISAND 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 semi-parametric efficiency bound in the homoskedastic regression model with $E[\varepsilon_i|z_i] = 0$ (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 (1) 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 non-zero 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 $\varepsilon_i|z_i$ is right-skewed for $z_i \geq 0$ and left-skewed for $z_i < 0$, this downweighing tends to occur mostly with positive disturbances when $z_i \geq 0$, and negative disturbances if $z_i < 0$, which leads to a negative bias in the estimation of β.

The much larger risk IMIX relative to ISAND (or INLR) under DCAS suggests that one must be quite sure of the statistical independence of ε_i and z_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.[4] Alternatively, one might of course model a potential conditional asymmetry of $\varepsilon_i|z_i$, although I am not aware of such an approach in the Bayesian literature in the context of a linear regression.[5]

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 pseudo-true value

[4]Under DCAS, a Bayesian model selection or averaging over IMIX and INLR would closely approximate IMIX, since the mixture model reduces the Kullback-Leibler divergence (1) by about 0.02 (so that with $n = 200$, the odds ratio has expected value of about $e^4 \simeq 55$). One might argue that visual inspection of the residuals would easily reveal the misspecification in the conditional shape of $\varepsilon_i|x_i$; but DGPs with similar effects as DCAS can also be constructed for higher dimensional regressors where such misspecifications are harder to diagnose.

[5]There are well developed Bayesian approaches for modelling heteroskedasticity through a scalar multiplication of the disturbance by some function of the regressors—see, for instance, Leslie, Kohn, and Nott (2007) for a recent contribution and references. But a more general model along those lines would not improve the bias of β in DCAS, since in DCAS, the variance of ε_i does not vary as a function of z_i. For computational reasons, we did not include heteroskedastic Bayesian models in our numerical comparison, as the estimation of risk requires a large number of posterior simulations.

remains consistent whenever the disturbances are mean independent of the regressors. Further, as predicted by Theorem 2, replacing the posterior of this model by the artificial 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 Empirical Illustration

As an empirical illustration, consider Kose, Otrok and Whiteman's (2003) study of international business cycles, using a panel of output, consumption and investment growth in 60 countries in yearly data for the years 1961-1991. Specifically, they estimate the following model for the demeaned data series $y_{t,j}$

$$y_{j,t} = b_j^{\text{world}} f_t^{\text{world}} + b_j^{\text{region}} f_{r(j),t}^{\text{region}} + b_j^{\text{country}} f_{c(j),t}^{\text{country}} + u_{j,t},\ t = 1, \cdots, 30, \quad j = 1, \cdots, 180 \quad (25)$$

where f_t^{world}, $f_{r,t}^{\text{region}}$, $r = 1, \cdots, 7$ and $f_{c,t}^{\text{country}}$, $c = 1, \cdots, 60$, are unobserved global, regional and country specified factors, $r(j)$ and $c(j)$ are the region and country of series j, b_j are unobserved factor loadings and $u_{j,t}$ is a series specific idiosyncratic error term. The idiosyncratic error terms $u_{j,t}$ are stationary Gaussian autoregressive (AR) processes of order 3. Kose, Otrok, and Whiteman (2003) conduct Bayesian inference in model (25), with an independent standard normal prior for the factor loadings b_j, each set of three AR parameters are independent mean-zero multivariate Gaussian with covariance matrix $\text{diag}(1, 1/2, 1/4)$ (restricted to values that imply the largest root of the lag polynomial to be bigger than 1.001 in modulus), the driving disturbance in the AR(3) for the $u_{j,t}$'s are inverse Gamma with parameters (6,0.001), and the factors are independent draws from a stationary Gaussian AR(3) with the same prior on the AR parameters as the idiosyncratic shocks. For identification of the scale and sign of the factors, the driving disturbance in the factor autoregressions are fixed, b_j^{world} is non-negative for the US output series, b_j^{region} is positive for one output series in each region, and b_j^{country} is positive for the output series in each country. We refer to Kose, Otrok, and Whiteman (2003) for a more detailed description of the model and the data.

The choice of Gaussian errors in the factors and errors is computationally convenient. What is more, conditional on the factor loadings, the multivariate Gaussian nature of the model implies that factors are identified through the second moments of the data. Accordingly, Kose, Otrok, and Whiteman (2003) describe their estimation procedure on

page 1221 as a "decomposition of the second moment properties of the data (e.g., the spectral density matrix)". This decomposition is arguably of interest whether or not the series are multivariate Gaussian, leading to a pseudo-true interpretation of the factors. At the same time, by sufficiency, also the posterior *variances* from a multivariate Gaussian model depend on the data only through the first two moments. For instance, when a bivariate Gaussian model is fitted to pairs of independent observations, the asymptotic posterior variance for the correlation ρ is given by $(1-\rho^2)^2/(1+\rho^2)$. But this description for the sample uncertainty is only adequate when the relationship between the two variables is linear and homoskedastic. In non-linear or heteroskedastic models, correlations and covariances still describe the strength of linear associations, but their sample uncertainty is adequately described by the sample variance of the appropriate moment condition, which involves the fourth moment of the data. This suggests that it is useful to base inference in model (25) on the sandwich posterior. We focus in the following on inference about the world factor $\{f_t^{\text{world}}\}_{t=1}^{30}$.

As Kose, Otrok, and Whiteman (2003) note, since model (25) contains over 1600 parameters, direct application of maximum likelihood is not an attractive option. We therefore implement the algorithm of Section 2.4, where the 30×1 vector θ is the world factor, and γ collects all other unknowns. We consider two schemes for the revelation of the information in all of the data $X_n = \{\{y_{j,t}\}_{t=1}^{30}, j = 1, \cdots, 180\}$. In the first scheme (CROSS), the data of one country is added one by one to the information set—formally, in the notation of Section 2.4, $x_i = \{\{y_{j,t}\}_{t=1}^{30} : c(j) = i\}$, $i = 1, \cdots 60$, with the the countries ordered as in the list in the appendix of Kose, Otrok, and Whiteman (2003). The second scheme (TIME) treats the countries symmetrically and instead reveals more and more information about the 30 observations in time. It would seem natural to reveal one time period at the time. But such a scheme would concentrate most information about each element f_t^{world} of θ in the single step where $\{y_{j,t}\}_{j=1}^{180}$ is revealed. The resulting score sequence is thus not suitable for estimating sandwich matrix. For this reason, the scheme TIME reveals at each step an additional trigonometrically weighted average of the 180 series, starting with the low frequencies, $x_i = \{\sum_{t=1}^{30} \cos(\pi i(t-1/2)/30) y_{j,t},\ j = 1, \cdots, 180\}$, $i = 1, \cdots, 30$. The two schemes have a different focus for the potential form of misspecification beyond the functional form of the marginal distributions of the disturbances: Dynamic misspecification of model (25) will in general lead to correlated scores under TIME, while spatial misspecification (say, some regions have two factors rather than one) will lead to correlated

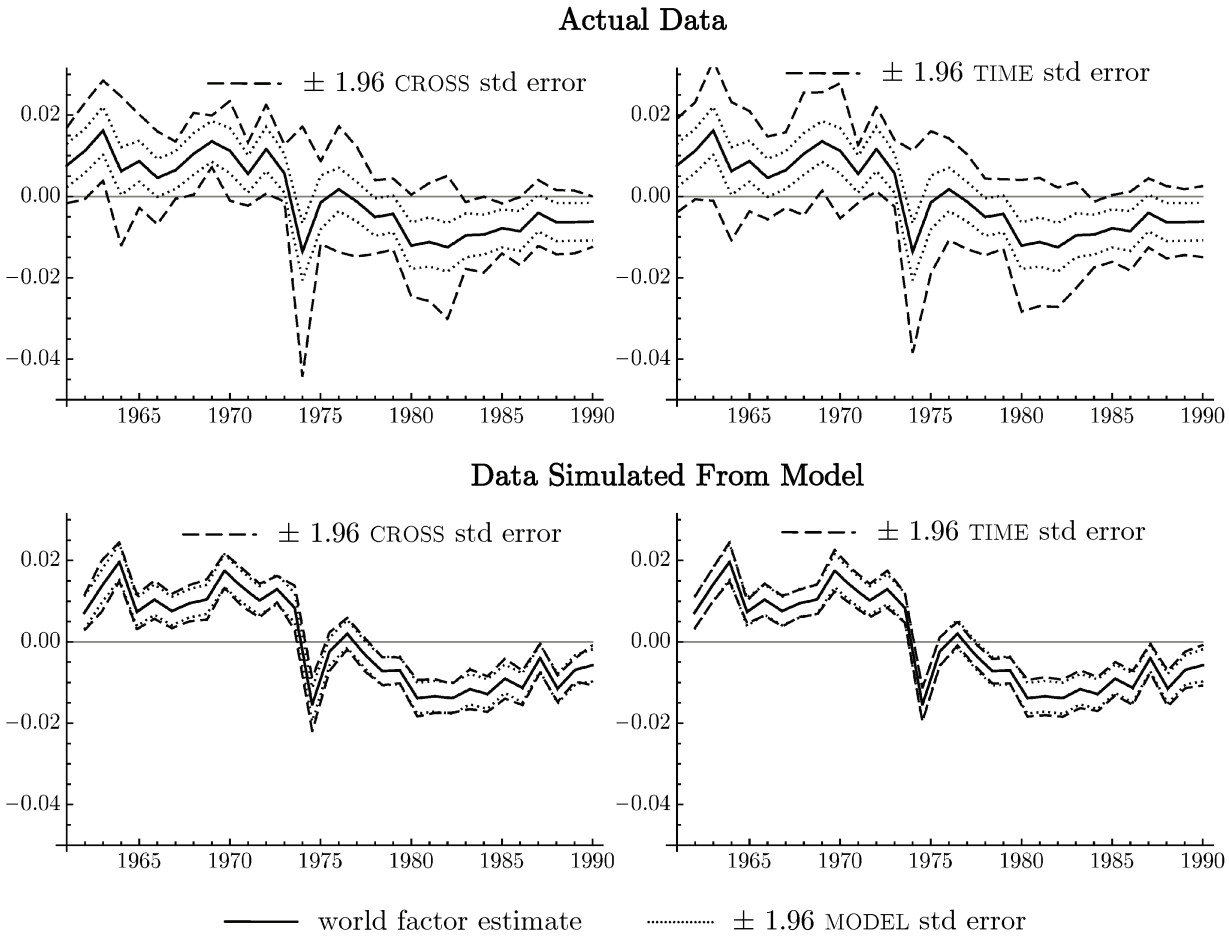
scores under the CROSS scheme.

The upper panel in Figure 2 depicts the estimate $\hat{\theta}$ of the world factor and three 95% interval estimates: one based on the posterior variance $\hat{\Sigma}_M/n$ of θ of the model (MODEL)[6], and two based on the sandwich covariance matrix estimators $\hat{\Sigma}_S/n$ for the two revelation schemes CROSS and TIME. For almost all elements of θ, the latter are substantially wider and often include the value zero, suggesting that the world factor is a more elusive concept than suggested by an analysis that assumes an entirely correct specification of the model. To get some sense for the accuracy of the approximations underlying the sandwich covariance matrix estimators, we simulated data from model (25) with $\theta = \hat{\theta}$ and γ drawn from the prior distribution. As can be seen in the lower panel of Figure 2, in this artificial data set from a correctly specified model, the estimators $\hat{\Sigma}_S$ and $\hat{\Sigma}_M$ yield very similar results, in accordance to the theory developed here.

6 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 and with the usual sandwich covariance matrix. The sandwich covariance matrix correction, which is routinely applied for the construction of confidence regions in frequentist analyses, thus has potentially an important role also in Bayesian studies of potentially misspecified models.

[6]The results for the posterior mean and model standard deviations are qualitatively similar, but not identical to what is reported in Kose, Otrok, and Whiteman (2003). The posterior sampler underlying Figure 2 seems to pass Geweke's (2004) joint distribution test.



Notes: Results are based on 100'000 draws from a Gibbs sampler. For the full model and the CROSS scheme, the sampler is the same as the one described in Kose, Otrok and Whiteman (2003), except that to improve mixing of the chain, the factors are drawn jointly from their conditional distribution given the other parameters. (A draw from this joint distribution can be implemented without inverting matrices larger than 30×30 by repeatedly applying the formula for partitioned inverses.) In the TIME scheme, unknown linear combinations of the data are treated as latent variables that add one more conditionally Gaussian step to the Gibbs sampler.

Figure 2: Model and Sandwich Standard Errors for the World Factor in Kose, Otrok and Whiteman's (2003) Model

7 Appendix

The following Lemma is used in the proof of Theorem 2.

Lemma 1 *If Σ_n, $n \geq 0$ is a sequence of stochastic matrices that are almost surely elements of $\mathcal{P}^k$ 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)). ■

Proof of Theorem 2:

(i) For any d_n, define $r_n^i(\theta_0, d_n) = E[\ell^i(0, d_n(X_n))]$, where here and below, the expectation is taken relative to P_{n,θ_0}. Note that $|r_n^i(\theta_0, d_n) - r_n(\theta_0, d_n)| \leq \sup_{a\in\mathcal{A}} |\ell_n(\theta_0, a) - \ell_n^i(0, a)| \to 0$ by Condition 3 (i), so if suffices to show the claim for $r_n^i(\theta_0, d_n)$. Define $\tilde{\ell}_n(u, a) = \ell_n(\theta_0 + u/\sqrt{n}, a)$, $\tilde{\ell}_n^i(u, a) = \ell_n^i(u/\sqrt{n}, a)$, $\hat{u}_n = \sqrt{n}(\hat{\theta} - \theta_0)$, $\Sigma_{S0} = \Sigma_S(\theta_0)$, $\Sigma_{M0} = \Sigma_M(\theta_0)$, 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), $\hat{\delta}_n = d_{TV}(\tilde{\Pi}_n, \mathcal{N}(0, \Sigma_M)) \xrightarrow{p} 0$. 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 $X_n = D_n(\omega, \theta_0)$, $D_n : \Omega \times \Theta \mapsto \mathbb{R}^{r\times n}$. 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^+.

By definition (17) and $(\hat{u}_n^+, X_n^+) \sim (\hat{u}_n, X_n)$,

$$\inf_{a\in\mathcal{A}} \int \tilde{\ell}_n(u + \hat{u}_n^+, a) d\tilde{\Pi}_n^+(u) = \int \tilde{\ell}_n(u + \hat{u}_n^+, d_{Mn}(X_n^+)) d\tilde{\Pi}_n^+(u) \tag{26}$$

P^+-almost surely. Also, by Condition 3 (ii), $\int \tilde{\ell}_n^i(u, a_n^*(\Sigma_{M0}^+))\phi_{\Sigma_{M0}^+}(u)du \leq \int \tilde{\ell}_n^i(u, \hat{a}_n(X_n^+))\phi_{\Sigma_{M0}^+}(u)du = \int \tilde{\ell}_n^i(u + \hat{u}_n^+, q(\hat{u}_n^+, \hat{a}_n(X_n^+)))\phi_{\Sigma_{M0}^+}(u)du$ for $\hat{a}_n(X_n^+) = q(-\hat{u}_n^+, d_{Mn}(X_n^+))$ almost surely for large enough n. Thus

$$\begin{aligned}
0 \leq & \int \tilde{\ell}_n^i(u, \hat{a}_n(X_n^+))\phi_{\Sigma_{M0}^+}(u)du - \int \tilde{\ell}_n^i(u, a_n^*(\Sigma_{M0}^+))\phi_{\Sigma_{M0}^+}(u)du \\
\leq & \int (\tilde{\ell}_n^i(u, \hat{a}_n(X_n^+)) - \tilde{\ell}_n(u + \hat{u}_n^+, d_{Mn}(X_n^+))\phi_{\Sigma_{M0}^+}(u)du \\
& - \int (\tilde{\ell}_n^i(u, a_n^*(\Sigma_{M0}^+)) - \tilde{\ell}_n(u + \hat{u}_n^+, q(\hat{u}_n^+, a_n^*(\Sigma_{M0}^+)))\phi_{\Sigma_{M0}^+}(u)du \\
& + \int \tilde{\ell}_n(u + \hat{u}_n^+, d_{Mn}(X_n^+))(\phi_{\Sigma_{M0}^+}(u)du - d\tilde{\Pi}_n^+(u)) \\
& - \int \tilde{\ell}_n(u + \hat{u}_n^+, q(\hat{u}_n^+, a_n^*(\Sigma_{M0}^+))))(\phi_{\Sigma_{M0}^+}(u)du - d\tilde{\Pi}_n^+(u))
\end{aligned}$$

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

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

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

$$\int (\tilde{\ell}_n^i(u + \hat{u}_n^+, a_n) - \tilde{\ell}_n(u + \hat{u}_n^+, a_n))\phi_{\Sigma_{M0}^+}(u)du \to 0$$

P^+-almost surely by dominated convergence using Conditions 2 and 3 (i). Thus, for P^+-almost all $\omega^+ \in \Omega^+$,

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

and $\hat{u}_n^+(\omega^+) \to \Sigma_{S0}^+(\omega^+)^{1/2}Z^+(\omega^+)$. Condition 3 (iii) therefore implies that also

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

for P^+-almost all $\omega^+ \in \Omega^+$. As almost sure convergence and $\tilde{\ell}_n^i \leq \bar{\ell}$ implies 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 $E^+[\tilde{\ell}_n^i(-\hat{u}_n^+, \hat{a}_n(X_n^+))] - E[\int \tilde{\ell}_n^i(u, a_n^*(\Sigma_{M0}))\phi_{\Sigma_{S0}}(u)du] \to 0$. But this implies via $r_n^i(\theta_0, d_{Mn}(X_n)) = E[\tilde{\ell}_n^i(0, d_{Mn}(X_n))] = E[\tilde{\ell}_n^i(-\hat{u}_n, q(-\hat{u}_n, d_{Mn}(X_n)))] = E^+[\tilde{\ell}_n^i(-\hat{u}_n^+, \hat{a}_n(X_n^+))]$ 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 non-stochastic element of $\mathcal{P}^k$ in the support of $\Sigma_J(\theta_0)$.

Then, similar to the proof of part (i), with $\hat{\theta}_n^+ = \theta_0 + \hat{u}_n^+/\sqrt{n}$, from (17) and Condition 4 (ii),

$$\begin{aligned}
0 \quad \leq \quad & \int \ell_n(\theta_0 + (u + \hat{u}_n^+)/\sqrt{n}, d_{Mn}(X_n^+))\phi_{\Sigma_M^+(\hat{\theta}_n^+)}du \\
& \qquad - \int \ell_n(\theta_0 + (u + \hat{u}_n^+)/\sqrt{n}, d_{Mn}^*(\hat{\theta}_n^+, \Sigma_M^+(\hat{\theta}_n^+)))\phi_{\Sigma_M^+(\hat{\theta}_n^+)}du \\
\leq \quad & \int \ell_n(\theta_0 + (u + \hat{u}_n^+)/\sqrt{n}, d_{Mn}(X_n^+))(\phi_{\Sigma_M^+(\hat{\theta}_n^+)}du - d\tilde{\Pi}_n^+(u)) \\
& \qquad + \int \ell_n(\theta_0 + (u + \hat{u}_n^+)/\sqrt{n}, d_{Mn}^*(\hat{\theta}_n^+, \Sigma_M^+(\hat{\theta}_n^+)))(d\tilde{\Pi}_n^+(u) - \phi_{\Sigma_M^+(\hat{\theta}_n^+)}du) \to 0
\end{aligned}$$

P^+-almost surely, since $d_{TV}(\mathcal{N}(0, \Sigma_{M0n}^+), \mathcal{N}(0, \Sigma_{M0}^+)) \to 0$ P^+-almost surely from Lemma 1 via $\Sigma_{M0n}^+ \to \Sigma_{M0}^+$ P^+-almost surely, and $\hat{\delta}_n^+ = d_{TV}(\tilde{\Pi}_n^+, \mathcal{N}(0, \Sigma_{M0n}^+)) \to 0$ P^+-almost surely by construction. Thus, by Condition 4 (iii), also $\ell_n(\theta_0, d_{Mn}(X_n^+)) - \ell_n(\theta_0, d_n^*(\theta_0 + (\Sigma_{S0}^+)^{1/2}Z^+/\sqrt{n}, \Sigma_J(\theta_0 + (\Sigma_{S0}^+)^{1/2}Z^+/\sqrt{n}))) \to 0$ P^+-almost surely, so that from $Z^+ \sim \mathcal{N}(0, I_k)$ independent of $(\Sigma_{S0}^+, \Sigma_{M0}^+, \Sigma_S^+(\cdot), \Sigma_M^+(\cdot)) \sim (\Sigma_S(\theta_0), \Sigma_M(\theta_0), \Sigma_S(\cdot), \Sigma_M(\cdot))$,

$$r_n(\theta_0, d_{Mn}) - E[\int \ell_n(\theta_0, d_{Mn}^*(\theta_0 + u/\sqrt{n}, \Sigma_M(\theta_0 + u/\sqrt{n}))\phi_{\Sigma_S(\theta_0)}(u)du] \to 0. \tag{27}$$

Since this argument can be invoked for η-almost all θ_0, (27) holds for η-almost all θ_0.

Pick a large $K > 0$, and define $\mathcal{B} = \{\theta \in \mathbb{R}^k : ||\Sigma_M(\theta)|| < K \text{ and } ||\Sigma_M(\theta)^{-1}||^{-1} < K\}$, $\dot{\ell}_n(\theta, a) = \mathbf{1}[\theta \in \mathcal{B}]\ell_n(\theta, a)$ 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 (27), dominated convergence, Fubini's Theorem and a change of variables

$$\begin{aligned}
&\int \dot{r}_n(\theta_0, d_{Mn})\dot{\eta}(\theta_0)d\theta_0 \\
= \;& E\int\int \dot{\ell}_n(\theta_0, d^*_{Mn}(\theta_0 + u/\sqrt{n}, \Sigma_M(\theta_0 + u/\sqrt{n})))\phi_{\Sigma_S(\theta_0)}(u)du\ \dot{\eta}(\theta_0)d\theta_0 + o(1) \qquad (28)\\
= \;& E\int\int \dot{\ell}_n(\theta + u/\sqrt{n}, d^*_{Mn}(\theta, \Sigma_M(\theta)))\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\dot{\eta}(\theta + u/\sqrt{n})du\ d\theta + o(1).
\end{aligned}$$

Now condition on 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 $K_2 > 0$, consider

$$\begin{aligned}
&\int_{||u||\le K_2} \dot{\phi}_{\Sigma_S(\theta+u/\sqrt{n})}(u)du \\
\ge \;& (2\pi)^{-k/2}\int_{||u||\le K_2} \mathbf{1}[\theta + u/\sqrt{n} \in \mathcal{B}][\inf_{||v||\le K_2} \det(\Sigma_S(\theta + v/\sqrt{n}))^{-1/2}] \\
&\qquad\qquad \cdot \exp[-\tfrac{1}{2}\sup_{||v||\le K_2} u'\Sigma_S(\theta + v/\sqrt{n})^{-1}u]du \\
\to \;& \int_{||u||\le K_2} \phi_{\Sigma_S(\theta)}(u)du
\end{aligned}$$

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

$$\int_{||u||>K_2} \dot{\phi}_{\Sigma_S(\theta+u/\sqrt{n})}(u)du \le \int_{||u||>K_2} K^{k/2}\exp[-\tfrac{1}{2}||u||^2K^{-1}]du$$

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)). Because $\dot{\eta}$ and $1/\dot{\eta}$ are bounded, also $\int (\dot{\eta}(\theta+u/\sqrt{n})/\dot{\eta}(\theta))\dot{\phi}_{\Sigma_S(\theta+u/\sqrt{n})}(u)du \to \int \phi_{\Sigma_S(\theta)}(u)du$ by dominated convergence, and $\dot{\eta}(\theta + u/\sqrt{n})/\dot{\eta}(\theta) \to 1$ again also implies $\int |(\dot{\eta}(\theta + u/\sqrt{n})/\dot{\eta}(\theta))\dot{\phi}_{\Sigma_S(\theta+u/\sqrt{n})}(u) - \phi_{\Sigma_S(\theta)}(u)|du \to 0$. Thus,

$$\begin{aligned}
&\int \dot{\ell}_n(\theta + u/\sqrt{n}, d^*_{Mn}(\theta, \Sigma_M(\theta)))\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\frac{\dot{\eta}(\theta + u/\sqrt{n})}{\dot{\eta}(\theta)}du \\
&\qquad - \int \dot{\ell}_n(\theta + u/\sqrt{n}, d^*_{Mn}(\theta, \Sigma_M(\theta)))\phi_{\Sigma_S(\theta)}(u)du \to 0
\end{aligned}$$

and since these convergences hold for $\dot{\eta}$-almost all θ, and both $\dot{\eta}$ and $1/\dot{\eta}$ are bounded,

$$\int\int \dot{\ell}_n(\theta + u/\sqrt{n}, d^*_{Mn}(\theta, \Sigma_M(\theta)))\phi_{\Sigma_S(\theta+u/\sqrt{n})}(u)\dot{\eta}(\theta + u/\sqrt{n})du\ d\theta \tag{29}$$
$$- \int\int \dot{\ell}_n(\theta + u/\sqrt{n}, d^*_{Mn}(\theta, \Sigma_M(\theta)))\phi_{\Sigma_S(\theta)}(u)du\ \dot{\eta}(\theta)d\theta \to 0$$

by dominated convergence. Since this convergence hold conditionally for almost all $(\Sigma_M(\cdot), \Sigma_S(\cdot))$, and the second term in (29) as well as (28) are bounded, they also hold unconditionally by dominated convergence, and the result follows.

The second claim follows analogously, using $d_{TV}(\mathcal{N}(0, \Sigma_S(\hat{\theta})), \mathcal{N}(0, \hat{\Sigma}_S))) \xrightarrow{p} 0$ under P_{n,θ_0} for η-almost θ_0.

Proof of Theorem 3:

By straightforward arguments assumption (iv) implies that the maximum likelihood estimator $\hat{\theta}_m$ is consistent, $\hat{\theta}_m \xrightarrow{p} \theta_0$. Thus, there exists a real sequence $k'_n \to 0$ such that $E\mathcal{T} \geq 1 - k'_n$ where $\mathcal{T} = \mathbf{1}[||\hat{\theta}_m - \theta_0|| < k'_n]$. From now on, assume is n large enough so that $\{\theta : ||\theta - \theta_0|| < k'_n\} \subset \Theta_0$. By condition (ii) and a Taylor expansion

$$\begin{aligned} 0 &= \mathcal{T} n^{-1/2} S_n(\hat{\theta}_m) \\ &= \mathcal{T} n^{-1/2} S_n(\theta_0) + \mathcal{T}\left(n^{-1}\int_0^1 H_n(\theta_0 + \lambda(\hat{\theta}_m - \theta_0))d\lambda\right) n^{1/2}(\hat{\theta}_m - \theta_0) \end{aligned} \tag{30}$$

where $H_n(\theta) = \sum_{i=1}^n h_i(\theta)$, and derivatives of the log-likelihood outside Θ_0 are defined to be zero. By assumption (v), $\mathcal{T}(n^{-1}H_n(\theta) + \Sigma_{M0}^{-1}) \xrightarrow{p} 0$, where $\Sigma_{M0} = \Sigma_M(\theta_0)$ so that the weak convergence in Condition 1 (i) for $\hat{\theta} = \hat{\theta}_m$ follows from 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

$$\mathcal{T} s_i(\hat{\theta}_m) = \mathcal{T} s_i(\theta_0) + \mathcal{T}\left(\int_0^1 h_i(\theta_0 + \lambda(\hat{\theta}_m - \theta_0))d\lambda\right)(\hat{\theta}_m - \theta_0)$$

we find

$$\mathcal{T}||n^{-1}\sum_{i=1}^n s_i(\hat{\theta}_m)s_i(\hat{\theta}_m)' - n^{-1}\sum_{i=1}^n s_i(\theta_0)s_i(\theta_0)'|| \leq 2\mathcal{T} n^{1/2}||\hat{\theta}_m - \theta_0|| \cdot (\sup_{i\leq n} n^{-1/2}||s_i(\theta_0)||)$$
$$\cdot \sup_{||\theta-\theta_0||<k'_n} n^{-1}\sum_{i=1}^n ||h_i(\theta)|| + \mathcal{T} n||\hat{\theta}_m - \theta_0||^2 \cdot \left(\sup_{||\theta-\theta_0||<k_n} n^{-1}\sum_{i=1}^n ||h_i(\theta)||\right)^2$$

and $n^{-1}\sum_{i=1}^n s_i(\hat{\theta}_m)s_i(\hat{\theta}_m)' \xrightarrow{p} V(\theta_0)$ 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

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

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

$$|\hat{a}_n - a_n| \leq \int \left| p(\theta_0 + n^{-1/2}u)\,\mathrm{LR}_n(u) - \pi(\theta_0)\widehat{\mathrm{LR}}_n(u) \right| du \tag{31}$$

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

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

For any n, let m_n be the smallest m such that simultaneously, $n > \sup_{m' \leq m} n^*(m')$, $n^{1/2}K(m^{-1}) > 1$ 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

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

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

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

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

$$\int |p(\theta_0 + n^{-1/2}u)\mathcal{M}_n(u)\,\mathrm{LR}_n(u) - \hat{p}\widehat{\mathrm{LR}}_n(u)|du \leq \int |p(\theta_0 + n^{-1/2}u) - \hat{p}|\mathcal{M}_n(u)\,\mathrm{LR}_n(u)du + \hat{p}\zeta_n$$

and

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

Since $\hat{a}_n = O_p(1)$ as shown above, and $p(\theta)$ is continuous at θ_0 by assumption (i), it suffices to show that $\zeta_n \xrightarrow{p} 0$ to obtain $d_{TV}(\Pi_n, \mathcal{N}(\hat{\theta}_m, \Sigma_{M0}/n)) \xrightarrow{p} 0$.

Define $\delta_n = n^{-1/2} S_n(\theta_0) - \Sigma_{M0}^{-1}\hat{u}$ and $\Delta_n(u) = n^{-1}\int_0^1 H_n(\theta_0 + \lambda n^{-1/2}u)d\lambda + \Sigma_{M0}^{-1}$, so that so that for all n large enough to ensure $\{\theta : ||\theta - \theta_0|| < m_n^{-1}\} \subset \Theta_0$, $\sup_{u\in\mathbb{R}^k} \mathcal{M}_n(u)|\operatorname{LR}_n(u)/\widehat{\operatorname{LR}}_n(u) - \exp[\delta_n' u + \frac{1}{2}u'\Delta_n(u)u]| = 0$. Thus, by Jensen's inequality

$$\begin{aligned}(2\pi)^{-k/2}|\Sigma_{M0}|^{-1/2}\zeta_n &= \int |1 - \mathcal{M}_n(u)\exp[\delta_n' u + \tfrac{1}{2}u'\Delta_n(u)u]|\phi_{\Sigma_{M0}}(u-\hat{u})du \\ &\leq \left(\int (1 - \mathcal{M}_n(u)\exp[\delta_n' u + \tfrac{1}{2}u'\Delta_n(u)u])^2\phi_{\Sigma_{M0}}(u-\hat{u})du\right)^{1/2}. \quad (32)\end{aligned}$$

By assumption (v),

$$\mathcal{M}_n(u)||\Delta_n(u)|| \leq c_n = \sup_{||\theta-\theta_0||\leq m_n^{-1}} n^{-1}||H_n(\theta) - H_n(\theta_0)|| + ||n^{-1}H_n(\theta_0) + \Sigma_{M0}^{-1}|| \xrightarrow{p} 0.$$

and

$$\begin{aligned}\int \mathcal{M}_n(u)\exp[2\delta_n' u + u'\Delta_n(u)u]\phi_{\Sigma_{M0}}(u-\hat{u})du &\leq \int \exp[2\delta_n' u + c_n u'u]\phi_{\Sigma_{M0}}(u-\hat{u})du \\ \int \mathcal{M}_n(u)\exp[\delta_n' u + \tfrac{1}{2}u'\Delta_n(u)u]\phi_{\Sigma_{M0}}(u-\hat{u})du &\geq \int \exp[\delta_n' u - \tfrac{1}{2}c_n u'u]\phi_{\Sigma_{M0}}(u-\hat{u})du \\ &\quad - \int (1 - \mathcal{M}_n(u))\exp[\delta_n' u - \tfrac{1}{2}c_n u'u]\phi_{\Sigma_{M0}}(u-\hat{u})du.\end{aligned}$$

From (30) and assumption (v), $\delta_n \xrightarrow{p} 0$, 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 the Cauchy-Schwarz inequality,

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

and the convergence follows from $\int(1-\mathcal{M}_n(u))\phi_{\Sigma_{M0}}(u-\hat{u})du = \int_{||u||\geq n^{1/2}m_n^{-1}} \phi_{\Sigma_{M0}}(u-\hat{u})du \xrightarrow{p} 0$ and the same arguments as above. Thus, the r.h.s. of (32) converges in probability to zero, and $\zeta_n \geq 0$, so that $\zeta_n \xrightarrow{p} 0$.

Thus, $d_{TV}(\Pi_n, \mathcal{N}(\hat{\theta}_m, \Sigma_{M0}/n)) \xrightarrow{p} 0$, which implies that the posterior median $\hat{\theta}_\Pi$ satisfies $n^{1/2}(\hat{\theta}_m - \hat{\theta}_\Pi) \xrightarrow{p} 0$, and $n^{-1}\sum_{i=1}^n s_i(\hat{\theta}_\Pi)s_i(\hat{\theta}_\Pi)' \xrightarrow{p} V(\theta_0)$ follows from the same arguments used for $\hat{\theta} = \hat{\theta}_m$ above. Finally, $d_{TV}(\Pi_n, \mathcal{N}(\hat{\theta}_m, \Sigma_{M0}/n)) \xrightarrow{p} 0$ also implies that the posterior asymptotic variance of Π_n converges in probability to Σ_{M0}.

References

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

Bayarri, M. J., and J. O. Berger (1997): "Measures of Surprise in Bayesian Analysis," *Duke University working paper 97-46*.

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edn.

Blackwell, D. (1985): "Approximate Normality of Large Products," *Technical Report No. 54, Departement of Statistics, Berkeley*.

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

Bunke, O., and X. Milhaud (1998): "Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models," *Annals of Statistics*, 26, 617–644.

Chen, C. (1985): "On Asymptotic Normality of Limiting Density Functions with Bayesian Implications," *Journal of the Royal Statistical Society Series B*, 47, 540–546.

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

Domowitz, I., and H. White (1982): "Misspecified Models with Dependent Observations," *Journal of Econometrics*, 20, 35–50.

Dudley, R. M. (2002): *Real Analysis and Probability*. Cambridge University Press, Camebridge, UK.

Fahrmeir, L., and G. Tutz (2001): *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

Freedman, D. A. (2006): "Ont the so-Called "Huber Sandwich Estimator" and "robust Standard Errors"," *The American Statistician*, 60, 299–302.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004): *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edn.

Gelman, A., X. Meng, and H. Stern (1996): "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies," *Statistica Sinica*, 6, 733–807.

Geweke, J. (2004): "Getting It Right: Joint Distribution Tests of Posterior Simulators," *Journal of the American Statistical Association*, 99, 799–804.

——— (2005): *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

Hall, P., and C. C. Heyde (1980): *Martingale Limit Theory and its Applications*. Academic Press, New York.

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*, pp. 221–233, Berkeley. University of California Press.

Kleijn, B. J. K., and A. W. van der Vaart (2008): "The Bernstein-Von-Mises Theorem under Misspecification," *Working paper, Free University Amsterdam*.

Kose, M. A., C. Otrok, and C. H. Whiteman (2003): "International Business Cycles: World, Region, and Country-Specific Factors," *American Economic Review*, 93, 1216–1239.

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

Lancaster, T. (2003): "A Note on Bootstraps and Robustness," *Working paper, Brown University*.

Leslie, D. S., R. Kohn, and D. J. Nott (2007): "A General Approach to Heteroskedastic Linear Regression," *Statistics and Computing*, 17, 131–146.

Lieb, E. H., and M. Loss (2001): *Analysis*. American Mathematical Society, Providence, RI, 2nd edn.

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

Phillips, P. C. B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.

Royall, R., and T. Tsou (2003): "Interpreting Statistical Evidence by Using Imperfect Models: Robust Adjusted Likelihood Functions," *Journal of the Royal Statistical Society Series B*, 65, 391–404.

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

Schervish, M. J. (1995): *Theory of Statistics*. Springer, New York.

Sims, C. A., and H. Uhlig (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59, 1591–1599.

Stafford, J. E. (1996): "A Robust Adjustment of the Profile Likelihood," *Annals of Statistics*, 24, 336–352.

Szpiro, A. A., K. M. Rice, and T. Lumley (2007): "Model-Robust Bayesian Regression and the Sandwich Estimator," *UW Biostatistics Working Paper Series Nr. 320, University of Washington*.

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

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

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