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

Ulrich K. Müller
Princeton University

December 2015

Introduction

- Frequentist inference using MLE typically relies on "sandwich" covariance matrix to account for potential misspecification.
- No such correction in Bayesian inference.
- This paper
  - Relies on standard results in frequentist and Bayesian asymptotics for misspecified models.
  - Decision theoretic justification of sandwich correction: reduces risk (=loss averaged over random samples) of Bayesian inference about pseudo-true parameters in misspecified models.

$\Rightarrow$ Motivates use of sandwich correction also in Bayesian applied work.

Model and Pseudo-True Values

- x_i, $i = 1, \cdots, n$ i.i.d. sample with density $f(x)$.
- Fitted model has density $g(x, \theta)$, $\theta \in \Theta$. If $f(x) \neq g(x, \theta)$ for all $\theta \in \Theta \subset \mathbb{R}^k$, then $g(x, \theta)$ is misspecified.
- MLE $\hat{\theta}$ is consistent for pseudo-true value $\theta_0 = \arg\max_\theta E[g(x_i, \theta)]$, and

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_S),\ \Sigma_S = \Sigma_M V \Sigma_M$$

  where $\Sigma_M = E[-h_i(\theta_0)]^{-1}$ and $V = E[s_i(\theta_0)s_i(\theta_0)']$, with s_i and h_i the score and Hessian, respectively.
- Example: Linear regression with coefficient θ, fitted model $g(x, \theta)$ assumes normal errors independent of regressors, true model $f(x)$ has non-normal, possibly heteroskedastic errors. Pseudo-true value θ_0 is population regression coefficient, Σ_S are White standard errors.
- Important assumption: if model is misspecified, seek good inference about pseudo-true value θ_0.

Bayesian Asymptotics

- Posterior density $\Pi_n(\theta)$ is proportional to product of prior density $p(\theta)$, and likelihood $\exp[L_n(\theta)] = \exp[\sum_{i=1}^n \ln g(x_i, \theta)]$,

$$\Pi_n(\theta) \propto p(\theta) \exp[L_n(\theta)].$$

- By second order Taylor expansion, for all fixed $u \in \mathbb{R}^k$

$$\begin{aligned} L_n(\hat{\theta} + n^{-1/2}u) - L_n(\hat{\theta}) &= n^{-1/2}u' \sum_{i=1}^n s_i(\hat{\theta}) + \tfrac{1}{2}u' \left(n^{-1} \sum_{i=1}^n h_i(\tilde{\theta}) \right) u \\ &\xrightarrow{p} -\tfrac{1}{2}u'\Sigma_M^{-1}u \end{aligned}$$

  so that

$$L_n(\theta) \approx C - \tfrac{1}{2}n(\theta - \hat{\theta})'\hat{\Sigma}_M^{-1}(\theta - \hat{\theta}).$$

  (Bernstein-von Mises Theorem under misspecification, cf. Chen (1985), Bunke (1998), Gelman et al. (2004), Geweke (2005).)

Sandwich Likelihood

- Large sample shape of likelihood is *as if* one had observed

$$\hat{\theta} \sim \mathcal{N}(\theta, \Sigma_M/n).$$

  But, as noted above, actual sampling distribution of $\hat{\theta}$ is approximately

$$\hat{\theta} \sim \mathcal{N}(\theta, \Sigma_S/n),$$

  with $\Sigma_S \neq \Sigma_M$ in general.

- Suggests alternative by basing inference on "sandwich" log-likelihood L_{Sn} from model $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma_S/n)$,

$$L_{Sn}(\theta) = C - \tfrac{1}{2} n (\theta - \hat{\theta})' \Sigma_S^{-1} (\theta - \hat{\theta}).$$

- Study large sample risk (=average performance over repeated samples) of Bayesian inference using either original likelihood L_n, or L_{Sn}.

Gaussian Location Problem

- Large sample inference is just like in the $k \times 1$ model

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

with Σ_S known, where inference using the original likelihood mistakenly assumes $Y \sim \mathcal{N}(\theta, \Sigma_M/n)$, and inference using the sandwich likelihood uses the correct model.

- Denote by $\mathcal{A}$ the action space, by $\mathcal{D}$ the decision rules after observing Y (that is mappings $\mathbb{R}^k \mapsto \mathcal{A}$) and by ℓ the loss function $\ell : \Theta \times \mathcal{A} \mapsto [0, \infty)$.

- 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 ϕ_Σ is the density of $\mathcal{N}(0, \Sigma)$. Good decisions yield low risk.

Bayesian Inference Using Sandwich Likelihood

- Let $p(\theta)$ be a proper prior on Θ. The *Bayes risk* of decision d relative to the prior p equals

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

- The posterior density for θ is proportional to $\phi_{\mathbf{\Sigma}_S/n}(y-\theta)p(\theta)$. *Posterior expected loss* of action $a \in \mathcal{A}$ is thus proportional to

$$\int \ell(\theta, a)\phi_{\mathbf{\Sigma}_S/n}(y-\theta)p(\theta)d\theta \tag{1}$$

  and the Bayes's action after observing $Y = y$ minimizes (1).

- Note that

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

  so that the decision rule $d_S^* \in \mathcal{D}$ that minimizes posterior expected loss for each observation $Y = y$ also minimizes Bayes risk $R(p, d)$ over $d \in \mathcal{D}$.

Bayesian Inference Using Misspecified Likelihood

- The mistaken assumption $Y \sim \mathcal{N}(\theta, \mathbf{\Sigma}_M/n)$, $\mathbf{\Sigma}_M \neq \mathbf{\Sigma}_S$, will in general lead to the decision $d_M^* \in \mathcal{D}$ that minimizes, for each observation $Y = y$, posterior expected loss in the wrong model, i.e. $d_M^*(y)$ is the action that minimizes

$$\int \ell(\theta, a)\phi_{\mathbf{\Sigma}_M/n}(y - \theta)p(\theta)d\theta.$$

- By the optimality of d_S^*, $R(p, d_S^*) \leq R(p, d_M^*)$, and often, $R(p, d_S^*) < R(p, d_M^*)$.

- Estimation under squared loss example: Suppose $\Theta = \mathbb{R}$, $p = \mathcal{N}(0, \sigma_p^2)$, $A = \Theta$ and $\ell(\theta, a) = (\theta - a)^2$. Then, for $J = S, M$,

$$d_J^*(y) = \frac{\sigma_p^2 y}{\mathbf{\Sigma}_J/n + \sigma_p^2}$$

and a calculation shows $R(p, d_S^*) < R(p, d_M^*)$.

Dominating Sample Information and Invariance

- If n is large and p is continuous, then for $J = S, M$

$$\phi_{\Sigma_J/n}(y-\theta)p(\theta) \approx \phi_{\Sigma_J/n}(y-\theta)p(y)$$

and the posterior is approximately $\theta \sim \mathcal{N}(y, \Sigma_J/n)$.

- Loss function ℓ is invariant if for some $q : \Theta \times \mathcal{A} \mapsto \mathcal{A}$,

$$\ell(\theta_1, a) = \ell(\theta_2, q(\theta_1 - \theta_2, a)) \text{ for all } a \in \mathcal{A}, \theta_1, \theta_2 \in \Theta.$$

Example: estimation loss functions that depend only on $(a - \theta)$.

- Then
  - the Bayes decision rules d^*_S and d^*_M are invariant
  - for invariant rules $d^i(y)$, the frequentist risk $r(\theta, d^i)$ equals posterior expected loss $\int \ell(\theta, d^i(y))\phi_{\Sigma_S/n}(y-\theta)p(\theta)d\theta$

$\Rightarrow R(p, d^*_S) \leq R(p, d^*_M)$ is strengthened to $r(\theta, d^*_S) \leq r(\theta, d^*_M)$ for *all* $\theta \in \Theta$.

Summary So Far

- Large sample approximations for MLE and likelihood suggest that Bayesian inference based on original likelihood is just like Bayesian inference in the model $Y \sim \mathcal{N}(\theta, \mathbf{\Sigma}_S/n)$, but with the wrong assumption about the variance ($\mathbf{\Sigma}_M/n$ instead of $\mathbf{\Sigma}_S/n$). Inference based on sandwich likelihood corresponds to correct assumption about variance.

- A wrong variance assumption yields in general higher Bayes risk (=weighted average frequentist risk).

- Result is strengthened to higher frequentist risk for any true value of θ under invariant loss functions if the likelihood dominates the prior.

Formal Analysis

- **Main Condition:**

  (i) $\sqrt{n}\Sigma_S(\theta_0)^{-1/2}(\hat{\theta} - \theta_0) \Rightarrow Z \sim \mathcal{N}(0, I_k)$, and there exists an estimator $\hat{\Sigma}_S \xrightarrow{p} \Sigma_S(\theta_0)$, where $\Sigma_S(\theta_0)$ is a random psd $k \times k$ matrix independent of Z.

  (ii) $d_{TV}(\Pi_n, \mathcal{N}(\hat{\theta}, \Sigma_M(\theta_0)/n)) \xrightarrow{p} 0$, where $\Sigma_M(\theta_0)$ is a random psd $k \times k$ matrix independent of Z.

- **Additional Conditions:** (i) bounded loss; (ii) smooth loss function.

- **Main Result:** Large sample equivalence to Gaussian Location problem above, so that Bayesian inference with sandwich posterior $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma}_S/n)$ is uniformly lower risk than Bayesian inference based on misspecified likelihood.

- **Additional Result:** More primitive assumptions imply Main Condition, also for random $\Sigma_S(\theta_0)$ and $\Sigma_M(\theta_0)$.

Linear Regression Monte Carlo

- Model is

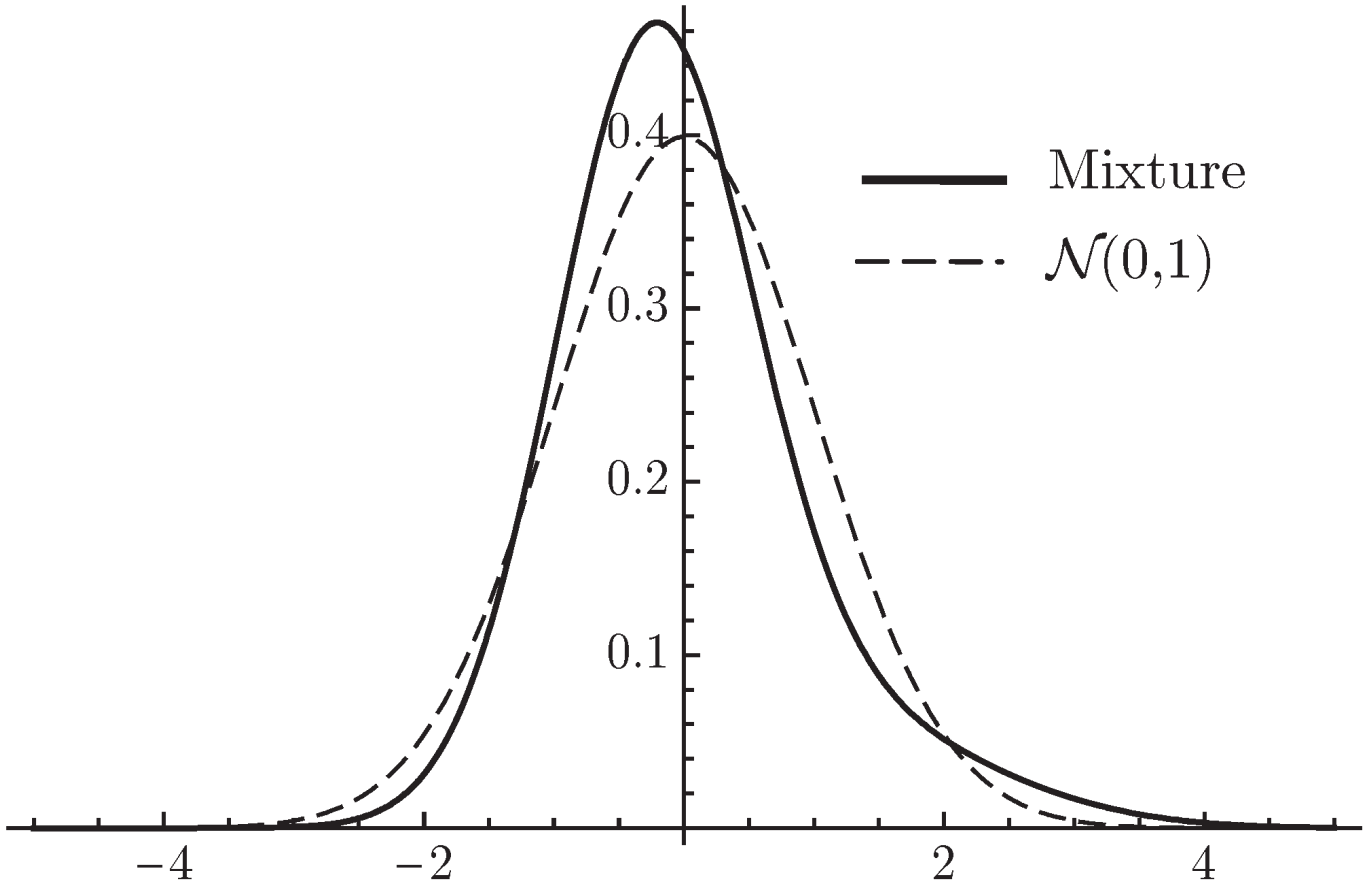
$$y_i = \alpha + x_i\beta + \varepsilon_i, \quad (y_i, x_i) \sim i.i.d., \quad i = 1, \cdots, n.$$

with $E[\varepsilon|x] = 0$, and $x \sim \mathcal{N}(0, 1)$.

Parameter is $\theta = (\alpha, \beta)$, and seek inference about population regression coefficient β.

- Data generating processes
  - DNLR: $\varepsilon|x \sim \mathcal{N}(0, 1)$
  - DMIX: $\varepsilon|x \sim \mathcal{M}$, where $\mathcal{M}$ is a right-skewed mean zero, unit variance mixture-of-two-normals
  - DCAS: $\varepsilon|x \sim (1 - 2 \cdot \mathbf{1}[x < 0])\mathcal{M}$
  - heteroskedastic versions of DNLR, DMIX and DCAS

Asymmetric Mixture-of-Two-Normals Density



Linear Regression Monte Carlo

- Modes of inference
  - INLR: normal linear regression model
  - IMIX: mixture of 3 normals with Dirichlet prior on mixing probabilities
  - ISAND: sandwich posterior in normal linear regression model (=OLS+White standard errors)
- (Bounded) Interval Estimation loss function with $a = (a_l, a_u) \in \mathbb{R}^2$, $a_l \leq a_u$

  $$\ell(\theta, a) = a_u - a_l + 40 \cdot \mathbf{1}[\beta < a_l](a_l - \beta) + 40 \cdot \mathbf{1}[\beta > a_u](\beta - a_u)$$

  Loss function rationalizes the usual 95% two-sided equal-tailed posterior probability interval.

Linear Regression Monte Carlo Results

Risk relative to ISAND

	homoskedasticity			heteroskedasticity		
	DNLR	DMIX	DCAS	DNLR	DMIX	DCAS
			$n = 50$			
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
			$n = 200$			
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
			$n = 800$			
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

Pseudo-true parameter under (IMIX, DCAS) is population regression coefficient minus 0.06.

Conclusion

- Combination of two standard results in frequentist and Bayesian asymptotics for misspecified models shows that risk of inference is lower if posterior action is computed for "sandwich" posterior.

- Sandwich covariance matrix has potentially important role also in Bayesian/decision theoretic analyses.