Bayesian Analysis (2009) **4**, Number 2, pp. 217–222

Rejoinder

Guosheng Yin*

I would like express my deep thanks to the Editor-in-Chief of *Bayesian Analysis*, Dr. Brad Carlin, for organizing these extensive discussions of my work on the Bayesian generalized method of moments (GMM). I am also grateful to the outstanding discussants: Drs. Ming-Hui Chen and Sungduk Kim, and Dr. Ciprian Crainiceanu for their insightful and stimulating comments on my article. In response to the suggestions from the discussants, I present some related computational issues, and also numerically examine the Bayesian GMM in the context of the least squares estimation and quantile regression. Many distributions are characterized by only the first and second moments, based on which the likelihood can be completely recovered. The Bayesian GMM is a robust, general and widely applicable approach, especially when there is not enough information to derive the likelihood.

1 Computational Issues

In the Bayesian GMM, the posterior distribution of the model parameters is quite complicated, and typically is not log-concave. Thus, I agree with Drs. Chen and Kim that the convergence of the Metropolis algorithm may be slow. In the frequentist GMM (Hansen 1982), $\widehat{\boldsymbol{\beta}}$ is computed via a two-stage iterative procedure by inserting the estimator obtained from the previous step, say $k-1$, into the covariance matrix, such that at the kth step, minimizing

$$Q_n^{(k)}(\boldsymbol{\beta}) = \mathbf{U}_n^T(\boldsymbol{\beta})\boldsymbol{\Sigma}_n^{-1}(\widehat{\boldsymbol{\beta}}^{(k-1)})\mathbf{U}_n(\boldsymbol{\beta})$$

with respect to $\boldsymbol{\beta}$ is much easier. A more efficient Markov chain Monte Carlo (MCMC) algorithm along the line of the two-stage iterative procedure is currently under development.

The Bayesian GMM is based on the moment conditions, instead of the likelihood, which can be applied as long as the moments are correctly specified. However, finding the correct moments is more difficult for longitudinal or clustered data because of the existing correlations, and is particularly challenging when the underlying correlation structure is complicated, as shown in the numerical studies by Drs. Chen and Kim. Although the Bayesian GMM with a working independence model is able to provide valid inferences, it may be less efficient, as in the case of the generalized estimating equation (GEE). Through personal communications with Drs. Chen and Kim, we agree that greater caution is required when selecting the appropriate moments. In some circumstances, the concatenated moments may have redundant information that would cause singularity of the covariance matrix. To resolve this, one can either delete the redundant rows or simply use the Moore-Penrose generalized inverse matrix when the

*Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, mailto:gsyin@mdanderson.org

© 2009 International Society for Bayesian Analysis DOI:10.1214/09-BA407REJ

covariance matrix is not of full rank.

Regarding Dr. Crainiceanu's comment on the motivation, the Bayesian and frequentist GMMs indeed produce quite similar results. However, when the dimensionality of $\boldsymbol{\beta}$ is high, it may be numerically difficult to minimize $Q_n(\boldsymbol{\beta})$ simultaneously over a large parameter space. The Bayesian GMM turns the multi-dimensional minimization problem into a series of one-dimensional Gibbs sampling procedures, and the posterior samples can be used for Bayesian inferences.

Dr. Crainiceanu also raised issues with the normalizing term, proper posterior distribution, and sampling the α_j's. The intuition behind the Bayesian GMM is that $Q_n(\boldsymbol{\beta})$ follows a chi-squared distribution, which means that it behaves exactly like $-2\log\{L(\mathbf{y}|\boldsymbol{\beta})\}$ in the usual likelihood ratio tests. Therefore, a pseudo-likelihood function $\widetilde{L}(\mathbf{y}|\boldsymbol{\beta})$ can be constructed as

$$\widetilde{L}(\mathbf{y}|\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}Q_n(\boldsymbol{\beta})\right\} = \exp\left\{-\frac{1}{2}\mathbf{U}_n^T(\boldsymbol{\beta})\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\beta})\mathbf{U}_n(\boldsymbol{\beta})\right\}. \tag{1}$$

By noting that the sample moment $\mathbf{U}_n(\boldsymbol{\beta})$ typically converges to a multivariate normal distribution, there is a normalizing term $(2\pi)^{-p/2}|\boldsymbol{\Sigma}_n(\boldsymbol{\beta})|^{-1/2}$. To address Dr. Crainiceanu's comment on the normalizing term and the associated computational issues, I conducted numerical studies in Sections 2 and 3, and found that it does not have much impact on the estimation as long as the kernel is in the form of (1).

Once the pseudo-likelihood $\widetilde{L}(\mathbf{y}|\boldsymbol{\beta})$ replaces the true likelihood function $L(\mathbf{y}|\boldsymbol{\beta})$, the normalizing constant for the posterior distribution is

$$c^{-1}(\mathbf{y}) = \int \exp\left\{-\frac{1}{2}\mathbf{U}_n^T(\boldsymbol{\beta})\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\beta})\mathbf{U}_n(\boldsymbol{\beta})\right\}\pi(\boldsymbol{\beta})d\boldsymbol{\beta},$$

where $\pi(\boldsymbol{\beta})$ is the prior distribution for $\boldsymbol{\beta}$. Because there exists a minimizer for $Q_n(\boldsymbol{\beta})$ under the regularity conditions (Hansen 1982), $\widetilde{L}(\mathbf{y}|\boldsymbol{\beta})$ has an upper bound so that $c(\boldsymbol{\beta})$ is finite. This leads to a proper posterior distribution for $\boldsymbol{\beta}$.

With correlated data, after splitting the moment conditions corresponding to each basis matrix $\mathbf{C}_{(j)}$, the α_j's drop out because the moments have a mean of zero:

$$\begin{aligned}
\mathbf{U}_{n(1)}(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i{}^T\mathbf{A}_i^{-1/2}\mathbf{C}_{(1)}\mathbf{A}_i^{-1/2}(\mathbf{y}_i-\boldsymbol{\mu}_i)\\
&\vdots\\
\mathbf{U}_{n(J)}(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i{}^T\mathbf{A}_i^{-1/2}\mathbf{C}_{(J)}\mathbf{A}_i^{-1/2}(\mathbf{y}_i-\boldsymbol{\mu}_i).
\end{aligned}$$

This is an example of overidentification, when there are more moments than the number of unknown parameters. This situation provides an opportunity to enhance the estimation efficiency, by concatenating the J moments, $\mathbf{U}_n(\beta) = \{\mathbf{U}_{n(1)}^T(\beta), \ldots, \mathbf{U}_{n(J)}^T(\beta)\}^T$.

2 Least Squares Estimation

In the usual linear regression model, $y_i = \boldsymbol{\beta}^T \mathbf{Z}_i + \epsilon_i$, where y_i is the outcome variable, $\mathbf{Z}_i$ is the covariate vector, and ϵ_i is the random error. The well-known least squares estimator (LSE) is obtained by minimizing the ℓ_2 norm $\sum_{i=1}^n (y_i - \boldsymbol{\beta}^T \mathbf{Z}_i)^2$. It can be cast in the maximum likelihood estimation under the assumption that $\epsilon_i \sim N(0, \sigma^2)$. However, the LSE does not require a normal distribution for the error, which simply minimizes the discrepancy between the observed values and the predicted values measured by the ℓ_2 norm, regardless of the distribution of the error. In the Bayesian GMM, the moment corresponding the LSE is

$$\mathbf{U}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i (y_i - \boldsymbol{\beta}^T \mathbf{Z}_i).$$

To examine the impact of different error distributions on the Bayesian GMM, I conducted simulation studies in the LSE framework. I considered the linear regression model,

$$y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \epsilon,$$

with the true parameter values $\beta_0 = 0.2$, $\beta_1 = 0.5$, $\beta_2 = -0.5$, and covariates $Z_1 \sim N(0, 1)$ and $Z_2 \sim \text{Bernoulli}(0.5)$. The error took five different distributions: $\epsilon \sim N(0, 0.25)$, $\epsilon \sim$ Cauchy or $t_{(1)}$, $\epsilon \sim t_{(3)}$, $\epsilon \sim$ Laplace, and a heteroscedastic error $\epsilon = \varepsilon Z_1$ with $\varepsilon \sim N(0, 0.25)$, respectively.

With sample sizes $n = 100$ and 200, I carried out 1,000 simulations. For each data replicate, 10,000 posterior samples were drawn for the Bayesian inference. For comparison, I also implemented the Bayesian GMM with the moment "normalizing constant" (nc=$|\boldsymbol{\Sigma}_n(\boldsymbol{\beta})|^{-1/2}$), and the usual Bayesian likelihood-based method assuming a normal error. I took noninformative prior distributions for all of the model parameters, i.e., $\beta_k \sim N(0, 10, 000)$, for $k = 0, 1, 2$; and $\sigma^{-2} \sim \text{Gamma}(0.0001, 0.0001)$.

The results in Table 1 show essentially no difference in the posterior estimates obtained from the two Bayesian GMMs, with or without the term nc=$|\boldsymbol{\Sigma}_n(\boldsymbol{\beta})|^{-1/2}$. When the error follows a normal distribution, the Bayesian likelihood method performs slightly better than the Bayesian GMM because the model is correctly specified. When the tails of the error distribution are heavier, such as in the Cauchy distribution, all of the methods do not produce consistent estimates. For the heteroscedastic error, i.e., the variance of the error is correlated with covariate Z_1, the coverage probability of β_1 is much lower when using the Bayesian likelihood approach. However, the Bayesian GMM does not rely on an assumption of homoscedastic errors, thus it still gives the correct variance estimate and allows the coverage probability to maintain the nominal level.

3 Quantile Regression

In contrast to the mean regression model, quantile regression is robust and gives an overall assessment of the covariate effects instead of examining only the central covariate

Table 1: Comparison between the Bayesian GMM, the Bayesian GMM with the normalizing constant $|\boldsymbol{\Sigma}_n(\boldsymbol{\beta})|^{-1/2}$ (w/nc), and the full Bayesian likelihood method with respect to the least squares estimation.

			$\beta_0 = 0.2$				$\beta_1 = 0.5$				$\beta_2 = -0.5$			
n	Bayesian	Error	Ave	ESD	ASD	CP	Ave	ESD	ASD	CP	Ave	ESD	ASD	CP
100	GMM	Normal	.199	.072	.074	95.6	.497	.051	.056	96.3	−.498	.103	.104	95.4
		$t_{(3)}$	.204	.227	.235	95.8	.511	.165	.177	97.6	−.508	.334	.337	95.0
		Cauchy	.395	1.401	1.574	98.6	.431	1.185	1.273	98.3	−.785	1.801	2.348	98.8
		Laplace	.214	.366	.390	96.1	.506	.283	.310	96.6	−.523	.504	.546	96.0
		Hetero	.196	.066	.073	97.3	.499	.080	.094	97.6	−.496	.094	.104	97.3
	GMM (w/nc)	Normal	.199	.071	.073	95.4	.497	.051	.054	96.1	−.498	.103	.103	95.3
		$t_{(3)}$	.205	.227	.232	95.5	.510	.164	.171	97.1	−.509	.332	.332	94.6
		Cauchy	.378	1.332	1.525	98.1	.448	1.093	1.199	98.1	−.760	1.719	2.281	98.5
		Laplace	.214	.367	.385	96.2	.506	.282	.301	96.2	−.522	.505	.540	95.6
		Hetero	.196	.067	.072	96.7	.499	.082	.088	95.6	−.496	.095	.102	96.6
	Likelihood (Normal)	Normal	.199	.071	.072	95.1	.497	.050	.051	95.4	−.498	.103	.102	95.1
		$t_{(3)}$	.204	.230	.235	95.1	.509	.168	.167	95.4	−.508	.333	.332	94.5
		Cauchy	.458	1.535	1.895	98.2	.387	1.288	1.526	95.4	−.769	1.733	2.325	98.5
		Laplace	.214	.371	.384	95.7	.506	.284	.288	94.9	−.522	.508	.537	95.2
		Hetero	.195	.069	.071	94.5	.499	.086	.050	74.6	−.496	.099	.101	95.0
200	GMM	Normal	.200	.049	.051	95.3	.499	.035	.037	95.5	−.499	.070	.072	95.4
		$t_{(3)}$	.202	.172	.170	95.4	.506	.120	.125	96.0	−.509	.238	.244	95.6
		Cauchy	.387	1.386	1.546	98.1	.451	1.114	1.206	98.2	−.761	1.858	2.316	98.8
		Laplace	.210	.275	.286	96.6	.513	.198	.209	95.9	−.512	.392	.404	95.4
		Hetero	.198	.049	.051	95.8	.501	.059	.063	97.2	−.498	.070	.072	94.3
	Likelihood (Normal)	Normal	.200	.049	.050	95.5	.499	.035	.036	94.2	−.499	.070	.071	95.2
		$t_{(3)}$	.202	.173	.172	95.2	.505	.122	.121	95.6	−.509	.239	.243	96.0
		Cauchy	.445	1.518	1.885	98.2	.432	1.264	1.511	96.1	−.746	1.815	2.315	98.5
		Laplace	.210	.276	.284	96.1	.513	.198	.202	95.7	−.512	.393	.400	95.2
		Hetero	.198	.050	.050	95.3	.501	.061	.036	72.4	−.498	.071	.071	94.4

Ave is the average of the posterior means over 1,000 simulations, ESD is the empirical standard deviation, ASD is the average of the posterior standard deviations, and CP(%) is the coverage probability of the 95% credible intervals.

Table 2: Comparison between the Bayesian GMM, the Bayesian GMM with the normalizing constant $|\boldsymbol{\Sigma}_n(\boldsymbol{\beta})|^{-1/2}$ (w/nc), and the full Bayesian likelihood method with respect to the median regression.

			$\beta_0 = 0.2$				$\beta_1 = 0.5$				$\beta_2 = -0.5$			
n	Bayesian	Error	Ave	ESD	ASD	CP	Ave	ESD	ASD	CP	Ave	ESD	ASD	CP
100	GMM	Normal	.201	.077	.094	97.6	.500	.055	.071	97.1	$-.501$	.112	.134	97.5
		$t_{(3)}$	.204	.186	.214	97.0	.496	.122	.158	98.5	$-.507$	.256	.305	97.5
		Cauchy	.190	.241	.278	96.1	.510	.185	.222	97.4	$-.503$	.341	.397	97.7
		Laplace	.192	.302	.370	97.2	.495	.242	.295	96.9	$-.476$	.436	.527	97.4
		Hetero	.200	.040	.047	96.1	.498	.078	.083	95.8	$-.500$	.056	.067	97.0
	GMM	Normal	.200	.077	.094	97.6	.500	.055	.071	97.3	$-.501$	.112	.135	97.6
	(w/nc)	$t_{(3)}$	.204	.186	.215	96.9	.496	.122	.158	98.5	$-.507$	.256	.307	97.5
		Cauchy	.191	.241	.280	96.6	.509	.185	.223	97.6	$-.503$	.341	.399	97.9
		Laplace	.192	.302	.372	97.5	.495	.243	.297	96.9	$-.475$	.436	.530	97.6
		Hetero	.200	.040	.047	96.0	.497	.077	.083	95.6	$-.501$	.056	.069	96.9
	Likelihood	Normal	.200	.069	.153	1	.500	.055	.119	1	$-.495$	.097	.213	1
	(Laplace)	$t_{(3)}$	.203	.185	.240	98.3	.495	.121	.169	99.5	$-.508$	.259	.341	98.9
		Cauchy	.193	.227	.266	98.1	.510	.169	.195	97.6	$-.503$	.324	.376	97.8
		Laplace	.191	.300	.308	95.7	.493	.228	.226	95.1	$-.478$	.428	.437	96.0
		Hetero	.201	.038	.121	1	.497	.076	.130	99.8	$-.499$	.053	.171	1
200	GMM	Normal	.200	.056	.066	97.1	.500	.040	.048	96.8	$-.499$	.080	.094	97.4
		$t_{(3)}$	.206	.129	.146	97.3	.501	.089	.106	97.6	$-.505$	.178	.207	97.6
		Cauchy	.201	.162	.181	96.4	.508	.115	.136	97.7	$-.501$	.225	.256	97.2
		Laplace	.193	.222	.247	96.5	.503	.154	.187	97.7	$-.477$	.316	.352	96.4
		Hetero	.200	.025	.027	94.8	.499	.054	.057	95.8	$-.500$	.035	.039	95.6
	Likelihood	Normal	.200	.055	.113	1	.501	.040	.081	1	$-.498$	.079	.160	1
	(Laplace)	$t_{(3)}$	.206	.131	.168	99.1	.501	.090	.119	99.5	$-.506$	.182	.237	98.7
		Cauchy	.201	.157	.184	97.7	.508	.111	.132	98.1	$-.502$	.220	.260	98.1
		Laplace	.195	.218	.213	94.7	.501	.150	.153	95.9	$-.481$	.307	.303	94.0
		Hetero	.200	.025	.079	1	.498	.054	.090	99.9	$-.500$	.036	.113	1

Ave is the average of the posterior means over 1,000 simulations, ESD is the empirical standard deviation, ASD is the average of the posterior standard deviations, and CP(%) is the coverage probability of the 95% credible intervals.

effects (Koenker and Bassett 1978). More importantly, quantile regression does not assume any distribution on the error, except for a conditional quantile of zero.

The τth $(0 < \tau < 1)$ quantile regression model takes the form of $q_\tau(y_i|\mathbf{Z}_i) = \boldsymbol{\beta}^T\mathbf{Z}_i$, where $q_\tau(y_i|\mathbf{Z}_i)$ is the conditional τth quantile given $\mathbf{Z}_i$, and $q_\tau(\epsilon_i|\mathbf{Z}_i) = 0$ with the distribution of the error ϵ_i unspecified. The estimator $\widehat{\boldsymbol{\beta}}$ can be obtained by minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{\beta}^T\mathbf{Z}_i), \tag{2}$$

where the check function $\rho_\tau(u) = u\{\tau - I(u < 0)\}$. Therefore, quantile regression is not a likelihood-based approach, and a Bayesian counterpart appears to be nonintuitive.

The check function is closely related to the asymmetric Laplace distribution (ALD) with a density function $f(y|\mu,\tau) = \tau(1-\tau)\exp\{-\rho_\tau(y-\mu)\}$, where μ is the location parameter. Minimizing (2) is equivalent to maximizing the likelihood function of y_i by assuming y_i from an ALD with $\mu = \boldsymbol{\beta}^T\mathbf{Z}_i$.

Focusing on the median regression with $\tau = 1/2$, I carried out simulation studies with the Bayesian GMM. In this case, the usual least absolute deviation (LAD) estimator is obtained by minimizing the ℓ_1 norm, $\sum_{i=1}^n |y_i - \boldsymbol{\beta}^T\mathbf{Z}_i|$, and the corresponding sample moment is

$$\mathbf{U}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i\{I(y_i - \boldsymbol{\beta}^T\mathbf{Z}_i \geq 0) - 1/2\}.$$

The simulation setups in the LAD are the same as those in the LSE, whereas the full Bayesian likelihood approach assumes an ALD error distribution.

The results, summarized in Table 2, indicate that the point estimates using both the Bayesian GMM and the Bayesian likelihood method are generally consistent. When the true error distribution is a Laplace distribution that exactly matches the model error assumption, the Bayesian likelihood method performs the best in terms of the variance estimate and the coverage probability. In other scenarios, the variances using the Bayesian likelihood approach are overestimated, which leads to coverage probabilities close to 1. In contrast, the Bayesian GMM reasonably maintains the coverage probabilities at the nominal level. With $n = 100$, the Bayesian GMM also inflates the variance estimates; as n increases to 200, the variance estimates improve. In summary, the Bayesian GMM is more robust as it produces reasonable estimates under various error distributions; whereas the Bayesian likelihood approach assuming a Laplace error appears to be quite sensitive to the model assumption.

References

Hansen, L. P. (1982). "Large sample properties of generalized method of moments estimators." *Econometrica*, 50: 1029–1054. 217, 218

Koenker, R. and Bassett, G. J. (1978). "Regression quantiles." *Econometrica*, 46: 33–50. 222