Getting it Right: Joint Distribution Tests of Posterior Simulators

John Geweke
University of Iowa

February, 2004
Forthcoming, *Journal of the American Statistical Association*

Abstract

Analytical or coding errors in posterior simulators can produce reasonable but incorrect approximations of posterior moments. This article develops simple tests of posterior simulators that detect both kinds of errors, and uses them to detect and correct errors in two previously published papers. The tests exploit the fact that a Bayesian model specifies the joint distribution of observables (data) and unobservables (parameters). There are two joint distribution simulators. The marginal-conditional simulator draws unobservables from the prior and then observables conditional on unobservables. The successive-conditional simulator alternates between the posterior simulator and an observables simulator. Formal comparison of moment approximations of the two simulators reveals existing analytical or coding errors in the posterior simulator.

Financial support from National Science Foundation grants SES-9996332 and SES-0214303 is gratefully acknowledged.

1 Introduction

In the past decade posterior simulators have become essential and widely used tools for Bayesian inference. In particular Markov chain Monte Carlo has made Bayesian inference routine in models that were previously inaccessible. As investigators developing these approaches are keenly aware, posterior simulators require analytic work for their proper implementation that can be difficult, tedious, or both. Data, prior and other densities must correspond exactly to models; conditional distributions must be derived correctly; and the computer code incorporating all of these ideas must be free of error. These tasks are essential, but there has been little attention given to formal verification that posterior simulators are error-free.

This paper proposes tests of the consistency of a posterior simulator with the specified prior and data distributions. These tests have power against errors of analysis and derivation, on the one hand, as well as failure to implement these ideas correctly in computer code, on the other. The tests utilize prior and data simulators that are natural complements of posterior simulators in statistical software. Given all three simulators clients and other second-party users can readily apply joint distribution tests to verify that a posterior simulator provided by an investigator is error-free.

2 Joint distribution tests

Let a model specify the distribution of a vector of observables $\mathbf{y} \in Y \subseteq \mathbb{R}^T$ conditional on a vector of unobservables $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$, by means of a density $p(\mathbf{y} \mid \boldsymbol{\theta})$ with respect to a measure μ on Y. Further let the model specify a proper prior distribution of unobservables by means of a density $p(\boldsymbol{\theta})$ with respect to a measure ν on Θ. The unobservables may be parameters, or a combination of parameters and latent variables, and the density $p(\boldsymbol{\theta})$ may incorporate a hierarchical structure. For the joint distribution tests it is essential that the prior distribution be proper, but by casting an improper prior distribution as a limiting special case of a proper prior distribution, joint distribution tests may be applied to posterior simulators for models with improper priors, as illustrated in Section 4.

In a model of this form

$$p(\boldsymbol{\theta}, \mathbf{y}) = p(\boldsymbol{\theta})\, p(\mathbf{y} \mid \boldsymbol{\theta}) \tag{1}$$

is the joint density of observables and unobservables, and (1) is a marginal-conditional decomposition of the joint density. Let g be any function $g : \Theta \times Y \rightarrow \mathbb{R}^1$ for which

$$\int_\Theta \int_Y g^2(\boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta}, \mathbf{y})\, d\mu(\mathbf{y})\, d\nu(\boldsymbol{\theta}) < \infty, \tag{2}$$

or more succinctly, $\sigma_g^2 = var\left[g(\boldsymbol{\theta}, \mathbf{y})\right] < \infty$.

A joint distribution test compares two simulation approximations of

$$\overline{g} = E\left[g(\boldsymbol{\theta}, \mathbf{y})\right] = \int_\Theta \int_Y g(\boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta}, \mathbf{y})\, d\mu(\mathbf{y})\, d\nu(\boldsymbol{\theta})$$

for a set of *test functions* g satisfying (2). The first approximation employs the *marginal-conditional simulator* of the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$,

$$\begin{aligned} \boldsymbol{\theta}^{(m)} &\sim p(\boldsymbol{\theta}), & (3)\\ \mathbf{y}^{(m)} &\sim p\left(\mathbf{y} \mid \boldsymbol{\theta}^{(m)}\right), & (4)\\ g^{(m)} &= g\left(\boldsymbol{\theta}^{(m)}, \mathbf{y}^{(m)}\right). \end{aligned}$$

This simulator is typically simple to construct, often much simpler than the posterior simulator. The sequence $\left\{\boldsymbol{\theta}^{(m)}, \mathbf{y}^{(m)}\right\}$ is i.i.d., $\overline{g}^{(M)} = M^{-1}\sum_{m=1}^M g^{(m)} \overset{a.s.}{\rightarrow} \overline{g}$, $M^{1/2}\left(\overline{g}^{(M)} - \overline{g}\right) \overset{d}{\rightarrow} N\left(0, \sigma_g^2\right)$, and $\widehat{\sigma}_g^{2(M)} = M^{-1}\sum_{m=1}^M \left(g^{(m)}\right)^2 - \left(\overline{g}^{(M)}\right)^2 \overset{a.s.}{\rightarrow} \sigma_g^2$.

In the posterior distribution of $\boldsymbol{\theta}$ the observable $\mathbf{y}$ is fixed at its observed (i.e., data) value $\mathbf{y}^o$. A posterior simulator produces a sequence of simulations $\left\{\widetilde{\boldsymbol{\theta}}_{\mathbf{y}^o}^{(m)}\right\}$, according to a transition kernel

$$\widetilde{\boldsymbol{\theta}}_{\mathbf{y}^o}^{(m)} \sim q\left(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}}_{\mathbf{y}^o}^{(m-1)}, \mathbf{y}^o\right). \tag{5}$$

Essentially all Markov chain Monte Carlo simulators can be expressed in this form. Careful application of these methods requires a demonstration that the sequence $\left\{\widetilde{\boldsymbol{\theta}}_{\mathbf{y}^o}^{(m)}\right\}$ is ergodic with unique invariant kernel $p(\boldsymbol{\theta})\, p(\mathbf{y}^o \mid \boldsymbol{\theta})$; [Tierney 1994] provides formal definitions and convergence conditions.

The *successive-conditional simulator* of the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$ consists of an initial draw $\widetilde{\boldsymbol{\theta}}^{(0)} \sim p(\boldsymbol{\theta})$, followed by the successive iterations

$$\widetilde{\mathbf{y}}^{(m)} \sim p\left(\mathbf{y} \mid \widetilde{\boldsymbol{\theta}}^{(m-1)}\right), \; \widetilde{\boldsymbol{\theta}}^{(m)} \sim q\left(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}}^{(m-1)}, \widetilde{\mathbf{y}}^{(m)}\right), \; \widetilde{g}^{(m)} = g\left(\widetilde{\boldsymbol{\theta}}^{(m)}, \widetilde{\mathbf{y}}^{(m)}\right).$$

If one has at hand a demonstration of the egodicity of $\left\{\widetilde{\boldsymbol{\theta}}_{\mathbf{y}^o}^{(m)}\right\}$ for almost all $\mathbf{y}^o$, then showing that $\left\{\widetilde{\boldsymbol{\theta}}^{(m)}, \widetilde{\mathbf{y}}^{(m)}\right\}$ is ergodic with unique invariant kernel $p(\boldsymbol{\theta}, \mathbf{y})$ typically involves little, if any, additional work. In this case $\widetilde{\overline{g}}^{(M)} = M^{-1} \sum_{m=1}^{M} g\left(\widetilde{\boldsymbol{\theta}}^{(m)}, \widetilde{\mathbf{y}}^{(m)}\right) \xrightarrow{a.s.} \overline{g}$. For a uniformly ergodic chain $\left\{\boldsymbol{\theta}^{(m)}, \mathbf{y}^{(m)}\right\}$, $M^{1/2}\left(\widetilde{\overline{g}}^{(M)} - \overline{g}\right) \xrightarrow{d} N\left(0, \tau_g^2\right)$.

If all simulators are error-free, then as $M_1 \to \infty$ and $M_2 \to \infty$,

$$\left(\overline{g}^{(M_1)} - \widetilde{\overline{g}}^{(M_2)}\right) / \left(M_1^{-1} \widehat{\sigma}_g^{2(M_1)} + M_2^{-1} \widehat{\tau}_g^{2(M_2)}\right)^{1/2} \xrightarrow{d} N(0,1). \tag{6}$$

The *Bayesian Analysis, Computation and Communication* software, freely available at `http://www2.cirano.qc.ca/~bacc`, computes both $\widehat{\tau}_g^{2(M)}$ and the test statistic (6).

Each test function g defines a joint distribution test (6). The power of the test depends on the error in the posterior simulator and the nature of the test function, and so a given error will become apparent with fewer iterations for some test functions than for other test functions. Using a wider variety of test functions provides a greater opportunity to detect existing errors in fewer iterations. Concerns about multiple tests can be addressed formally using a Bonferroni test of the joint hypothesis involving all test functions, as illustrated below in Section 4. In the examples in the following sections, errors, if they are present, emerge rather quickly in a subset of the test functions.

The marginal-conditional and successive-conditional simulators incorporate three simulators: the prior simulator (3), the observables simulator (4), and the posterior simulator (5). Thought processes and coding for these simulators are distinct. The observables simulator (4) appears in both the marginal-conditional and successive-conditional simulators, but an error in (4) will have different consequences for the marginal-conditional and successive-conditional simulators and lead to rejections in the joint distribution tests. While failure in the joint distribution tests can be due to errors in any of (3), (4) and (5), (5) is usually the most likely candidate since it is typically much more complicated than either (3) or (4).

3 A constructed example: *t*-mixture model

To demonstrate the kinds of errors joint distribution tests can detect, consider the univariate mixture of Student-t distributions

$$y_t \sim t\left(\mu_1, \sigma_1^2; \nu\right) \text{ with probability } p, \tag{7}$$

$$y_t \sim t\left(\mu_2, \sigma_2^2; \nu\right) \text{ with probability } 1-p. \tag{8}$$

In this example ν is fixed at $\nu = 5$, but the model could be extended to make ν an unknown parameter.

Standard MCMC algorithms [Geweke 1993] exploit the fact that the sequence $\nu\omega_t \sim \chi^2(\nu)$ followed by $y_t \sim N\left(\mu_j, \sigma_j^2/\omega_t\right)$ is equivalent to $y_t \sim t\left(\mu_j, \sigma_j^2; \nu\right)$ $(j = 1, 2)$. The model is augmented with $\boldsymbol{\omega} = (\omega_1, ..., \omega_T)'$ and the latent state vector $\mathbf{s} = (s_1, ..., s_T)'$, with $s_t = 1$ indicating (7) and $s_t = 2$ indicating (8). Then Gaussian priors for μ_1 and μ_2, inverse gamma priors for σ_1^2 and σ_2^2, and a beta prior for p are all conditionally conjugate, and the resulting conditional distributions in a Gibbs sampling algorithm are also Gaussian, inverse gamma, and beta, respectively. There are two variants of the Gibbs sampler. The first (MCMC1) draws $\mathbf{s}$ and $\boldsymbol{\omega}$ jointly and the second (MCMC2) draws them separately.

The marginal-conditional simulator draws the parameter vector $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)'$ from the prior and then simulates six observations $\mathbf{y} = (y_1, \ldots, y_6)'$ conditional on $\boldsymbol{\theta}$. The successive-conditional simulator alternates between the simulation of $\mathbf{y}$ conditional on the unobservables $\boldsymbol{\theta}$, $\mathbf{s}$ and $\boldsymbol{\omega}$, and an iteration of MCMC1 or MCMC2. In this illustration the test functions are the five first and fifteen second moments of the parameter vector $\boldsymbol{\theta}$. (Since $\mathbf{y}$ is not involved in the comparison, it really is not necessary to generate $\mathbf{y}^{(m)}$ in the marginal-conditional simulator.)

Five instances of an intentionally introduced error provide evidence on the power of the posterior simulation check.

1. The prior distribution of p is Beta(1,1), whereas the posterior simulator assumes a Beta(2,2) prior distribution.

2. In the successive-conditional simulator the observables simulator ignores $\boldsymbol{\omega}$ from the posterior simulator. Instead it uses fresh values of $\nu\omega_t \sim \chi^2(\nu)$ to construct y_t.

3. The variance of the Gaussian conditional posterior distribution of $(\mu_1, \mu_2)'$ is erroneously set to zero.

4. The degrees of freedom in the conditional posterior distribution of each ω_t is taken to be 5, rather than its correct value of 6.

5. The final error is in the generation of (s_t, ω_t) in MCMC1. The correct algorithm generates s_t (conditional on all unknowns except ω_t) and then generates ω_t conditional on all unknowns including s_t just drawn. In the error, ω_t is drawn several steps later in the Gibbs sampling algorithm rather than immediately after s_t.

Table 1: Summary of p-values of test statistics in the t-mixture model

		Tests (of 20) failing at $p =$			
Algorithm	Error	.05	.01	.005	.001
MCMC1	0. None	0	0	0	0
MCMC2	0. None	0	0	0	0
MCMC1	1. Prior simulation of p	4	3	3	2
MCMC1	2. Simulation of observables	10	9	9	9
MCMC1	3. ω degrees of freedom	5	3	3	3
MCMC1	4. μ variance	11	10	10	9
MCMC1	5. (s, ω) draw	7	6	6	6

NOTE: Tests compare approximations of 20 test functions: the five first and fifteen second moments of five parameters in the model. Errors are detailed in the text. The numbers of functions failing the test at alternative p-values are reported. Tests utilized 2.5×10^5 iterations of each simulator.

Table 1 reports the number of rejections in the twenty joint distribution tests, using some alternative conventional critical values. The joint distribution tests employed the twenty first and second moments of the 5×1 parameter vector $\boldsymbol{\theta}$ as test functions. Both simulators employed 2.5×10^5 iterations, and total computing time was less than one minute. The correct algorithm clearly passes the joint distribution tests, whereas errors—in the prior, observables, or posterior simulators—are all detected.

4 An example: Unit roots model

Errors in posterior simulators can lead to incorrect but reasonable results. If the errors are undetected then incorrect results are likely to be published. An example of such a published study is [Geweke 1994]. That article develops a variant of the univariate time series models that have been applied by economists in the past two decades to discriminate between trend and difference stationarity in macroeconomic time series.

The distribution of the observable time series $\{y_t\}$ in [Geweke 1994] is specified by

$$\begin{aligned} y_t &= \gamma + \delta t + u_t; && (9) \\ u_t &= \rho u_{t-1} + \sum_{j=1}^{m} a_j \left(u_{t-j} - u_{t-j-1}\right) + \varepsilon_t; && (10) \\ & \varepsilon_t \overset{i.i.d.}{\sim} t\left(0, \sigma^2; \nu\right). && (11) \end{aligned}$$

This and similar models have been applied widely, usually to natural logarithms of time series measuring output, employment, prices or money supply. The main parameter of interest is ρ. If $\rho \in (0,1)$ then $\{u_t\}$ is stationary and, from (9), $\{y_t\}$ is said to be trend stationary. If $\rho = 1$ then $\{u_t\}$ is nonstationary, it follows that $\Delta y_t = \delta\left(1 - \sum_{j=1}^{m} a_j\right) + \sum_{j=1}^{m} a_j \Delta y_{t-j} + \varepsilon_t$ and $\{y_t\}$ is said to be difference stationary. An advantage of the particular formulation (9)-(10) is that $E\left(\Delta y_t\right) = \delta$ in either case, which facilitates articulation of a prior distribution for the unconditional growth rate without conditioning on trend or difference stationarity, and also facilitates interpretation of the posterior distribution.

Conditional on difference stationarity, the prior distribution fixes $\rho = 1$. Conditional on trend stationarity, the prior density of ρ is

$$p(\rho) = (s+1)^{-1} \rho^s I_{(0,1)}(\rho). \qquad (12)$$

This prior distribution is motivated by two properties of the simple first-order autoregressive process $y_t = \rho y_{t-1} + v_t$. First, for observations at $(s+1)$-period intervals, $y_t = \rho_{(s)} y_{t-s} + v_t^{(s)}$, with $\rho_{(s)} = \rho^s$. Second, if $p\left(\rho_{(s)}\right) = I_{(0,1)}\left(\rho_{(s)}\right)$, then the prior density of ρ is (12). In this context a larger value of s in (12) corresponds to a "flat" prior for the autoregressive parameter $\rho_{(s)}$, and places relatively more weight on values of ρ near $\rho = 1$. The leading objective of the posterior analysis in [Geweke 1994] is to determine Bayes factors among trend stationary models with different specifications of s in (12), and between each of these models and a difference-stationary model ($\rho = 1$).

The prior distribution in [Geweke 1994] is completed by setting $m = 4$ on the basis of previous research by several investigators, and then specifying $a_j \sim N\left(0, .731 \cdot .342^j\right)$ so that $a_1 \sim N\left(0, .5^2\right)$ while $a_4 \sim N\left(0, .1^2\right)$. Other prior distributions are $\delta \sim N\left(0, .05^2\right)$, $\gamma \sim N\left(0, 10^2\right)$, $\nu \sim \exp(4)$ and

$$p(\sigma) \propto \sigma^{-1} I_{(0,\infty)}(\sigma). \qquad (13)$$

The latter improper reference prior can be interpreted as the limit of inverse gamma distributions

$$a/\sigma^2 \sim \chi^2(a) \text{ as } a \to 0. \qquad (14)$$

The study [Geweke 1994] shows that the Bayes factor in favor of the specification $s = t$ in (12), versus $s = r$, is the expectation of the function

$$g(\rho) = \left[(t+1)/(r+1)\right] \rho^{(t-r)}$$

under the posterior distribution corresponding to the prior distribution with $s = r$ in (12). In the context of the improper prior distribution (13), the interpretation of the Bayes factor is that of the limiting Bayes factor of models compared with a common value of a in (14). The Bayes factor in favor of the difference stationarity specification $\rho = 1$ is the expectation of the function

$$g\left(\rho\right) = \frac{\left(s+1\right)^{-1} \exp\left[-\left(1-\widehat{\rho}\right)^2/2\lambda^2\right]}{\int_0^1 \rho^s \exp\left[-\left(\rho-\widehat{\rho}\right)^2/2\lambda^2\right] d\rho}, \tag{15}$$

the parameter s being that of the prior density (12). The parameters $\widehat{\rho}$ and λ in (15) are the mean and standard deviation parameters in the conditional posterior distribution of ρ, which is normal truncated to the unit interval. The one-dimensional integral in the denominator is evaluated by conventional quadrature.

The posterior simulator developed in [Geweke 1994] is a Gibbs sampling algorithm utilizing the six blocks (γ, δ), $(a_1, \ldots, a_4)$, ρ, ν, σ, and a block for latent variables introduced to manage the Student-t distributions of the innovations (11) as described in the previous section. All six conditional distributions are derived (correctly, as it turns out) in [Geweke 1994], except for two typographical errors not incorporated in the computer code: the expression $N\left(0, \pi_0\pi_1^{j-1}\right)$ in the conditional distribution of $(a_1, \ldots, a_4)$ should read $N\left(0, \pi_0\pi_1^{j}\right)$, and the leading term ρ in the conditional kernel density of ρ, equation (18) in [Geweke 1994], should read ρ^s.

To conduct joint distribution tests, archived computer code was retrieved, and the results published in [Geweke 1994] were replicated exactly, using random number generator seeds that had been archived along with the code. (The only change at all was that execution time decreased by a factor of 100, consistent with Moore's law and the decade vintage difference in the workstations used for [Geweke 1994] and the joint distribution test.) A final step was added to each iteration of the Gibbs sampler to simulate the observables from (9)-(11). The prior distributions just described were used in the joint distribution tests, except that the improper prior (13) was replaced with the proper inverse gamma prior $.01/\sigma^2 \sim \chi^2\left(4\right)$.

Test functions involved only parameters, not observables, and therefore the marginal-conditional simulator involved only simulation from the prior distribution; 10^6 replications, requiring about 30 seconds, were used. The successive-conditional simulator employed $T = 10$ observations and $M = 4 \times 10^6$ iterations of which every 400'th iteration was recorded and used in the joint distribution tests, which required about 25 minutes. There was no evidence of serial correlation in the successive-conditional simulator at intervals of 400 iterations.

Ninety test functions were used in the joint distribution comparison tests: (a) all nine parameters of the model (ρ, γ, δ, $a_1, \ldots, a_4$, σ, and ν) (b) the three functions $f_1 = \rho^9$, $f_2 = \gamma\left(1-\rho\right) + \delta\sum_{j=1}^{4} a_j$ and $f_3 = \delta\left(1-\rho\right)$; and (c) all squares and cross-products of the twelve functions in (a) and (b). Table 2 provides the outcome of the tests. It indicates parameters and functions in its first row and column. The remaining entries correspond to products of the first row and column, entries being made only on and below the diagonal. Note that the test statistics in the last row correspond to the respective column headings and involve no interaction between parameters or functions. The entries are the absolute values of test statistics that all have standard normal distributions if all relevant algorithms have been derived and coded correctly. Only entries that exceed 1.96 in absolute value are shown. A Bonferroni test rejects the null hypothesis that the two simulators are the same at the 0.1% level if any test statistic exceeds $\Phi^{-1}\left[.001/\left(2\cdot 90\right)\right] = 4.394$. These 37 entries indicate categorically the existence of one or more errors in the algorithm or coding.

Table 2: Joint distribution test statistics in the unit roots model

	ρ	γ	δ	a_1	a_2	a_3	a_4	σ	ν	f_1	f_2	f_3
ρ	2.4											
γ	2.5											
δ			3.4									
a_1				19.5								
a_2	7.5	7.8			11.8							
a_3				2.0		11.7						
a_4						2.7	11.7					
σ	9.8	10.4			7.3			4.3				
ν	4.4	4.4			6.0			2.2	2.6			
f_1								5.9	3.6			
f_2	2.1	2.6			7.4			7.4	2.1		3.3	
f_3			3.3					2.3			2.1	3.3
1.0	2.6				7.9			10.4	4.3		2.6	

NOTE: The test function for any entry is the product of the row parameter or function and column parameter or function; functions are described in the text. Test statistics are based on every 400'th of 4×10^6 iterations. Test statistics are shown only if they exceed 1.96 in absolute value.

In general joint distribution tests indicate the existence of errors, but not their source. Nevertheless not much is lost by looking first at conditional distributions of parameters that generate the most egregious test failures. The main diagonal in Table 2 singles out $a_1, \ldots, a_4$, and if one adds to this group σ, then all test statistics exceeding 5.0 in Table 2 involve only these parameters. With this clue, inspection of the code quickly isolated one error: the routine employed to draw from the posterior conditional Gaussian distribution of $(a_1, \ldots, a_4)'$ in fact drew from the marginal distribution of $(a_1, \ldots, a_4)'$ in the joint conditional posterior distribution of $(a_1, \ldots, a_4, \sigma)'$. The same routine was used in the draw from the conditional posterior distribution of $(\gamma, \delta)'$ and the same error was made. The error was easily corrected. Repetition of the joint distribution comparison tests yielded only two rejections in tests of size .05 and none in tests of size .01.

Table 3: Comparison of Bayes factors and moments in error-ridden and corrected posterior MCMC simulators, time series model for real per capita GNP

Bayes factor or parameter	With error ($M = 10^4$)	With error ($M = 10^6$)	Corrected ($M = 10^6$)	Corrected ($M = 10^4$)
BF favoring $\rho = 1$	.526 [.026]	.445 [.004]	.0462 [.0004]	.0482 [.0026]
BF favoring $s = 17$	.722 [.018]	.639 [.002]	.274 [.001]	.290 [.011]
ρ	.896 (.068)	.886 (.071)	.849 (.063)	.852 (.062)
δ	.017 (.004)	.018 (.004)	.018 (.001)	.018 (.001)
ν	6.1 (3.8)	6.0 (3.8)	5.3 (3.6)	5.3 (3.5)

NOTE: The prior distribution of ρ takes $s = 9$ in (12). Numbers in brackets for Bayes factors are numerical standard errors. In the last three rows the simulation approximation of the posterior expectation of the indicated parameter is shown, together with the posterior standard deviation in parentheses.

In this context it is interesting to examine whether the detected error had significant consequences for the questions taken up in [Geweke 1994]. Table 3 provides some results of this examination for one of the six time series used in that study, per capital real GNP. The table pertains to a trend stationary model with specification of the prior distribution (12) for ρ in which $s = 9$. The first row of entries indicates the computed Bayes factor in favor of difference stationarity against this specification. The second row indicates the Bayes factor in favor of an alternative trend stationary model with greater persistence, $s = 17$ in the prior distribution for ρ in (12). In these rows the numbers in brackets indicate the numerical standard error (i.e., the standard error of the Monte Carlo approximation, correcting for serial correlation in the simulator output) corresponding to the numerical approximation of the Bayes factor. The last three rows indicate posterior means and (in parentheses) posterior standard deviations of three of the parameters.

Thanks to the increase in computing speed since the original study, it was practical to increase the number of Monte Carlo replications by a factor of 100, as indicated in the first row of Table 3. (Computing time was about 15 minutes on a 2000-vintage Hewlett-Packard workstation using compiled Fortran.) In the process of this replication, it was discovered that increasing the number of burn-in iterations from 200 to 2000 and the number of retained iterations from 10^4 to 10^6 in the original error-ridden code, itself significantly changed some posterior moments of interest. This is evident in the comparison of columns two and three in Table 3. However the impact of the incorrect coding on the Bayes factors, evident in the comparison of columns three and four in this table, is an order of magnitude greater. The Bayes factor in favor of difference stationarity was too high by a factor of ten, and the Bayes factor in favor of difference stationarity with $s = 17$ in (12) was also much too high. Consistent with these errors, the effect of the coding error was to render the posterior mean of ρ about 0.90 when it should have been about 0.85. (The impact of the coding error on Bayes factors and the posterior mean of ρ was in the same direction and of the same order of magnitude for the other time series studied in [Geweke 1994], as well.) There was little impact on inference about trend (γ), but the coding error lowered the degrees of freedom in the Student-t distribution of the innovations somewhat. The last column of Table 3 repeats the exercise with the corrected code, utilizing the same number of burn-in and retained iterations applied originally. The small differences in the last two columns of the table, compared with columns two and three, suggest that convergence to the invariant distribution is faster in the correct code than in the error-ridden code.

5 An Example: Reduced rank regression model

The study [Geweke 1996] is a second instance of a published study based on an incorrect posterior simulator. That article proposes posterior simulators for a Bayesian treatment of the reduced rank regression model, first introduced in [Anderson 1951] and used subsequently in such diverse applications as simultaneous equation estimation [Dreze 1976], inference for cointegrated time series [Johansen 1988] and asset pricing models [Costa et al. 1997].

To establish notation let

$$\underset{T\times L}{\mathbf{Y}} = \underset{T\times p}{\mathbf{X}} \cdot \underset{p\times L}{\mathbf{\Theta}} + \underset{T\times k}{\mathbf{Z}} \cdot \underset{k\times L}{\mathbf{A}} + \underset{T\times L}{\mathbf{E}} \tag{16}$$

denote the multivariate regression model for T observations of L dependent variables and $p+k$ covariates. Conditional on covariates,

$$vec\left(\mathbf{E}\right) \sim N\left(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}_T\right). \tag{17}$$

The reduced rank regression model is the special case in which

$$\underset{p\times L}{\mathbf{\Theta}} = \underset{p\times q}{\mathbf{\Psi}} \cdot \underset{q\times L}{\mathbf{\Phi}}, \;\; q < \min\left(p, L\right).$$

Further restrictions are required to identify $\mathbf{\Psi}$ and $\mathbf{\Phi}$, and [Geweke 1996] considers two. In normalization 1, $\mathbf{\Phi} = \left[\begin{array}{cc} \mathbf{I}_q & \mathbf{\Phi}^* \end{array}\right]$ and $\mathbf{\Psi}$ is unrestricted. In normalization 2, $\mathbf{\Psi}' = \left[\begin{array}{cc} \mathbf{I}_q & \mathbf{\Psi}^{*\prime} \end{array}\right]$ and $\mathbf{\Phi}$ is unrestricted.

For the independent prior distributions

$$\mathbf{\Sigma}^{-1} \sim W\left(\underline{\mathbf{S}}^{-1}, \underline{\nu}\right), \tag{18}$$

and either

$$\left[vec\left(\mathbf{\Psi}\right), vec\left(\mathbf{\Phi}^*\right)\right]' \sim N\left(\mathbf{0}, \tau^{-2}\mathbf{I}_{L(p+q)}\right), \tag{19}$$

in the case of normalization 1, or

$$\left[vec\left(\mathbf{\Psi}^*\right), vec\left(\mathbf{\Phi}\right)\right]' \sim N\left(\mathbf{0}, \tau^{-2}\mathbf{I}_{L(p+q)}\right), \tag{20}$$

in the case of normalization 2, [Geweke 1996] develops a Gibbs sampling algorithm for the model. The body of [Geweke 1996] presents conditional posterior distributions for (18)-(19) and (18)-(20), and the appendix of the article derives these distributions in detail for the limiting case $\underline{\mathbf{S}}^{-1} \to \mathbf{0}$, $\underline{\nu} \to 0$, $\tau \to 0$.

Before conducting joint distribution tests, archived computer code was retrieved and posterior moments from the original work were duplicated. Code for the first simulator of the tests was written and checked, and the conditional data distributions (16)-(17) were added to the posterior simulator. In all tests $L = 4$, $p = 3$, $k = 0$, $q = 2$, $\tau = 1$, $\underline{\mathbf{S}} = 2.5\mathbf{I}_4$, $\underline{\nu} = 8$, and $T = 6$. The first column of $\mathbf{X}$ was set to units and the remaining elements of $\mathbf{X}$ to independent standard normal random variables. The test functions were $\mathbf{\Psi}$ and $\mathbf{\Phi}^*$ (normalization 1) or $\mathbf{\Psi}^*$ and $\mathbf{\Phi}$ (normalization 2), $\mathbf{\Theta}$, $\mathbf{\Theta}'\mathbf{\Theta}$, $\mathbf{\Sigma}^{-1}$, and the ordered eigenvalues $\lambda_1 \geq \ldots \geq \lambda_4$ of $\mathbf{Y}'\mathbf{Y}$. The marginal-conditional simulation employed $M = 10^6$ i.i.d. iterations, requiring about 6.5 minutes of computing. The successive-conditional simulation used every 40'th iteration of a total of 4×10^6 iterations, and required about one hour of computing. Using correct code, there is little evidence of serial correlation in every 40'th iteration.

Table 4: Some selected moment approximations in the reduced rank model, normalization 2 (error-ridden code)

	Full distribution		Numerical accuracy	
Test function	Mean	Standard.	Numerical.	Relative numerical
		deviation	standard error	efficiency
ψ^*_{11}	-12.0	23.6	22.9	.001
ψ^*_{12}	-2.25	18.7	7.85	.006
ϕ_{22}	-.208	.028	.063	.196
ϕ_{23}	.195	.951	.030	.076
θ_{22}	-.208	.885	.028	.063
θ_{34}	-2.52	14.2	4.92	.008
$\sum_{j=1}^{3} \theta^2_{j1}$	362.	654.	338.	.004
$\sum_{j=1}^{3} \theta_{j2}\theta_{j1}$	89.3	342.	106.	.010
σ^{11}	3.27	.053	.061	.738
σ^{21}	-.034	1.14	.033	1.17
λ_1	13,112.	10,836.	9,352.	.001
λ_2	37.9	32.0	1.76	.333

NOTE: Moments are based on the first 1,000 iterations of the successive-conditional simulator, employing normalization 2 of the reduced-rank regression model described in the text. For each test function in the first column, the second and third columns show corresponding approximation of the mean and standard deviation; the fourth and fifth columns show the numerical standard error and relative numerical efficiency of the mean approximation.

The tests produced strikingly different outcomes for the two normalizations. The initial test for normalization 1 found that two of the 46 test functions rejected at size .05 and none at .01. For normalization 2 the successive-conditional simulator failed due to exponent overflow after several thousand iterations. The algorithm was restarted, halted at $M = 10^3$ iterations, and the approximations $\overline{\overline{g}}^{(M)}$ and some related moments were computed. Table 4 provides these moments for each of a dozen of the 46 test functions. The first column indicates the test function and the second reports the corresponding approximation $\overline{\overline{g}}^{(M)}$ of its expected value $\overline{g}$, using the successive-conditional simulator. The third column is the corresponding approximation of its standard deviation, $\left[M^{-1}\sum_{m=1}^{M}\left(\widetilde{g}^{(m)}-\overline{\overline{g}}^{(M)}\right)^2\right]^{1/2}$. The fourth column is the numerical standard error of the approximation, $\left(\widehat{\tau}_g^{2(M)}/M\right)^{1/2}$ in the notation of Section 2. The last column is the relative numerical efficiency of the approximation – the number of iterations that would be required in an i.i.d. simulator to the number required in the actual successive-conditional simulation, to achieve the same numerical standard error. The results contrast markedly with the successive-conditional simulation for normalization 1, for which relative numerical efficiencies were all above 0.1 and most were near 1.0. Results for $\mathbf{\Psi}^*$, and hence for the last row of $\mathbf{\Theta}$, $\mathbf{\Theta}'\mathbf{\Theta}$, and the eigenvalues of $\mathbf{Y}'\mathbf{Y}$, are especially poor; those for $\mathbf{\Phi}$ are somewhat better; and those for $\mathbf{\Sigma}^{-1}$ taken in isolation would be no cause for concern.

Table 5: Joint distribution test statistics in the reduced rank model (original error-ridden code)

Function	Test	Function	Test
ϕ_{22}	3.27	θ_{23}	2.54
ϕ_{13}	2.28	$\sum_{j=1}^{3}\theta_{j4}^2$	2.99
ϕ_{23}	2.54	σ^{43}	2.90
θ_{22}	3.27	λ_2	7.76
θ_{13}	2.28	λ_3	3.33

NOTE: Test statistics taken from the same 1,000 iterations of the successive conditional simulator as in Table 4, and 10^6 iterations of the marginal-conditional simulator.

Joint distribution tests were conducted using these 10^3 iterations from the successive-conditional simulation, together with the 10^6 iterations of the marginal-conditional simulator. The tests turned up rejections for ten of the 46 functions in tests of size .05, six of size .005 and two of size .001, as indicated in Table 5. The power of the tests is weakened by the substantial numerical standard errors (shown in Table 4) from the successive-conditional simulator, yet the joint distribution tests detect the problems.

Careful inspection of the algorithm and code first focused on $\mathbf{\Psi}^*$, because of its poor behavior documented in Table 4. This inspection turned up the fact that in the conditional distribution of $\mathbf{\Psi}^*$ the precision parameter τ was taken to be $\tau = 0$ due to a programming error that had occurred when the coding was originally extended from the case of an uninformative prior to the shrinkage prior (20). This error corresponds to a correct coding of a prior Gaussian "distribution" for $\mathbf{\Psi}^*$ with precision zero, that is, a situation in which the prior distribution and therefore the full distribution of parameters and observables does not exist. The difficulties in the successive-conditional simulator conveyed in Table 4 reflect the fact that no invariant full distribution exists. Consistent with this interpretation, when the error was corrected the second simulator exhibited no signs of divergence, and in 4×10^6 iterations, with functions recorded every 40'th iterations, displayed the same ideal relative numerical efficiencies as the successive-conditional simulator with normalization 1.

Table 6: Some joint distribution test results in the reduced rank model (original code, partially corrected)

Function	Test	Function	Test	Function	Test
ϕ_{14}	2.11	$\sum_{j=1}^{3} \theta_{j1}^2$	15.03	λ_1	16.50
ϕ_{34}	2.25	$\gamma_{j=1}^{3} \theta_{j2}^2$	13.79	λ_2	20.32
θ_{14}	2.11	$\sum_{j=1}^{3} \theta_{j3}^2$	12.61	λ_3	16.30
θ_{24}	2.25	$\sum_{j=1}^{3} \theta_{j4}^2$	11.40	λ_4	6.58

NOTE: Test statistics are based on on every 40'th of 4×10^6 iterations of the successive conditional simulator, and the same 10^6 iterations of the marginal-conditional simulator utilized in Tables 4 and 5

Nevertheless problems remained after correction of this error. As indicated in Table 6 there were twelve rejections of size .05 and eight of size .001. Further inspection of the algorithm and code traced the error to the conditional distributions of $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ in normalization 2. If $\tau = 0$ then the joint conditional distribution of $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ has a convenient and well-known form (see Appendix A.1 of [Geweke 1996]). This joint conditional distribution was retained when the code was originally modified to accommodate $\tau > 0$. The error was corrected by incorporating the respective, correct conditional distributions of $\mathbf{\Phi}$ and $\mathbf{\Sigma}$. Repetition of the joint distribution test was then successful, with two rejections in tests of size .05 and none in tests of size .01.

6 Summary and conclusions

This article provides tests of the correctness of posterior simulators, based on formal comparison of marginal-conditional and successive-conditional simulators of the joint distribution of observables and unobservables in a fully specified Bayesian model. Examples illustrate that these tests detect analytical and coding errors that produce incorrect but reasonable results and may well otherwise pass unnoticed. These joint distribution tests require only that the relevant prior and data simulators be made available along with a posterior simulator. These are natural complements in Bayesian statistical software. Joint distribution tests may readily be used by investigators (including students and authors) to demonstrate the reliability of posterior simulators to clients (including professors and editors), and clients may independently use them to verify the correctness of a posterior simulator. Their routine use should increase the productivity of Bayesian investigators generally.

References

[Anderson 1951] Anderson, T.W. (1951), "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions," *Annals of Mathematical Statistics,* 29, 813-828.

[Costa et al. 1997] Costa, M., G. Attilio and P. Paolo (1997), "A Reduced Rank Regression Approach to Tests of Asset Pricing," *Oxford Bulletin of Economics and Statistics,* 59, 163-181.

[Dreze 1976] Dreze, J.H. (1976), "Bayesian Limited Information Analysis of the Simultaneous Equation Model," *Econometrica,* 44, 1045-1075.

[Geweke 1991] Geweke, J. (1991), "Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments," in J.O. Berger, et al. (eds.), *Bayesian Statistics*, Vol. 4, pp. 169-194. Oxford: Oxford University Press.

[Geweke 1993] Geweke, J. (1993), "Bayesian Treatment of the Independent Student-*t* Linear Model," *Journal of Applied Econometrics,* 8, S19-S40.

[Geweke 1994] Geweke, J. (1994), "Priors for Macroeconomic Time Series and Their Application," *Econometric Theory,* 10, 609-632.

[Geweke 1996] Geweke, J. (1996), "Bayesian Reduced Rank Regression in Econometrics," *Journal of Econometrics,* 75, 121-146.

[Geweke 1999] Geweke, J. (1999), "Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication" (with discussion and rejoinder), *Econometric Reviews,* 18, 1-126.

[Johansen 1988] Johansen, S. (1988), "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control,* 12, 231-254.

[Tierney 1994] Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion and rejoinder), *Annals of Statistics,* 22, 1701-1762.