Automated Sensitivity Computations for MCMC Gibbs Output

Dan Zhu
dan.zhu@monash.edu

December 4, 2019

MONASH University

Overview

Working Papers

Automated Sensitivity Analysis for Bayesian Inference via Markov Chain Monte Carlo: Applications to Gibbs Sampling. L. Jacobi, M.S. Joshi and Dan Zhu. Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2984054`

How Sensitive are VAR Forecasts to Prior Hyperparameters? An Automated Sensitivity Analysis. J.Chan, L. Jacobi and Dan Zhu. Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3185915`

MCMC

The key idea behind Bayesian MCMC-based inference is the construction of a Markov Chain with a transition kernel, $p(\theta^g|\theta^{g-1}, \eta_0, Y)$,that has the posterior distribution as its limiting distribution.

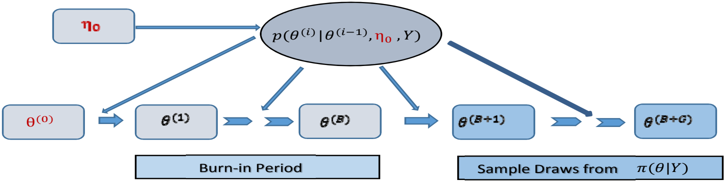


Figure: Algorithm based on MCMC chain.

Let θ_0 denote the vector $\{\eta_0, \theta^{(0)}\}$.

Sensitivity Analysis in MCMC

- Prior Robustness:

$$\frac{\partial \mathbb{E}_\pi[S(\boldsymbol{\theta})|Y, \boldsymbol{\eta}_0]}{\partial \boldsymbol{\eta}_0}$$

- Convergence: The choice of burning period

$$||\frac{\partial \boldsymbol{\theta}^{(g)}}{\partial \boldsymbol{\theta}^0}|| \leq \alpha$$

In many applications of simulation, we are primarily interest in computating

$$\alpha(\boldsymbol{\eta}_0) = \mathbb{E}_\pi[S(\boldsymbol{\theta})]$$

where $\boldsymbol{\theta} \sim \pi$. When π is known in full, independent sample are drawn. Derivatives with respect to the model inputs $\boldsymbol{\eta}_0$ is computed via three main methods.

Three main MC Derivative Estimation Method

The three traditional MC methods for derivative estimation:

1. Finite-Differencing Method(FD):
    - Computational cost
    - Unstable variance
2. Pathwise-Method(PW):
    - Dependent Sample
    - Discontinuous mapping
3. Likelihood-Ratio Method(LR):
    - Unstable variances
    - Only limited to $\frac{\partial \mathbb{E}_{\pi}[g(\boldsymbol{\theta})|Y,\boldsymbol{\eta}_0]}{\partial \boldsymbol{\eta}_0}$

Glasserman(2004) has detailed discussion of these three methods in the MC context.

The likeihood ratio: Perez et al(2006) and Müller 2015

Müller applies the approach described in Perez et al (2006) in the context of the exponential family to obtain the prior sensitivities of the β vector with respect to its prior mean vector $\mathbf{b}_0$,

$$\frac{\partial}{\partial \mathbf{b_0}} \mathbb{E}_{\hat{\pi}}[\beta|Y] = \Sigma_p^{-1} \Sigma_{\hat{\pi}}, \tag{1}$$

where Σ_p is the prior variance B_0 and $\Sigma_{\hat{\pi}}$ is the posterior covariance matrix of β.

Pathwise Method: the IPA Derivative

Typically, bayesians are interested on some sample statistics of the posterior distribution,

$$\alpha(\boldsymbol{\eta}_0) = \mathbb{E}_\pi[S(\boldsymbol{\theta})|Y, \boldsymbol{\eta}_0]$$

if we are interested in $\partial\alpha(\boldsymbol{\eta}_0)$, the pathwise or IPA derivative can be written as

$$\frac{1}{G}\sum_{g=B+1}^{G+B} J_S(\boldsymbol{\theta}^g)\frac{\partial\boldsymbol{\theta}^g}{\partial\boldsymbol{\eta}_0}.$$

Here $\boldsymbol{\theta}^g$ debnote both the gth draw as well as the mapping that samples it.

Gibbs Sampler

At each step of MCMC,

$$\boldsymbol{\theta}^g = \phi(\boldsymbol{\theta}^{g-1}, \boldsymbol{\eta}_0, \omega)$$

Hyper-parameter dependence

$$\frac{\partial \boldsymbol{\theta}^g}{\partial \boldsymbol{\eta}_0} = \frac{\partial \phi}{\partial \boldsymbol{\eta}_0} + \frac{\partial \phi}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}^{g-1}}{\partial \boldsymbol{\eta}_0}$$

Starting value dependence

$$\frac{\partial \boldsymbol{\theta}^g}{\partial \boldsymbol{\theta}^0} = \frac{\partial \phi}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}^{g-1}}{\partial \boldsymbol{\theta}^0}$$

Automatic Differentiation and the PW method

AD takes an algorithm for computing a value, E, and produces a new algorithm that computes the derivative of E with respect to its inputs.

- Computer program to evaluate a quantity then at its fundamental level, it is a string of elementary algebraic operations.
- An algorithm is just a *composite function* of these simple operations.

AD is the pathwise method of evolving the Jacobian matrix through the simple operations, via chain-rule!

Alternative Methods and Discontinuous mappings

For cases where F^{-1} does not exist or is too cumbersome to work with, alternative methods were introduced to simulate these variates. There are inherent discontinuities in these algorithms since a candidate outcome, x, is accepted as a variate from the target distribution if

$$a(x, \theta) \leq v \text{ for } v \sim U(0, 1).$$

Treatment for Gamma Variates

Glasserman and Liu (2010) proposed the distributional derivative method that obtains the derivatives of random variates X_θ with respect to its distributional parameters θ

$$\frac{\partial X_\theta}{\partial \theta} = -\frac{\frac{\partial}{\partial \theta} F(X, \theta)}{f(X, \theta)}.$$

We adapt the Glasserman and Liu(2010) method for computing the distributional derivatives the Gamma random variables and extended to treat Wishart random variates.

Quantile Sensitivities

Suppose our forecast random variable Y is absolutely continuous with the distribution $F_Y(\cdot;\boldsymbol{\theta}_0)$. For a given $\alpha \in (0,1)$, the α-quantile, denoted as Y^*, is defined implicitly by

$$F_Y(Y^*;\theta_0) = \alpha.$$

By the implicit function theorem, we have

$$\frac{\partial Y^*}{\partial \theta_0} = -\frac{\frac{\partial F_Y(y;\theta_0)}{\partial \theta_0}}{f_Y(y;\theta_0)}\Bigg|_{y=Y^*},$$

where $f_Y(\cdot;\theta_0)$ is the associated density function, which is unfortunately unknown.

However, suppose there exists a latent random vector $\mathbb{Z} \sim f_{\mathbb{Z}}(\cdot;\theta_0)$ such that

$$F_Y(y;\theta_0) = \mathbb{E}\left[G\left(y;\mathbb{Z}(\theta_0),\theta_0\right)\right]$$

for a function $G\left(y;\mathbb{Z}(\theta_0),\theta_0\right)$ that is absolutely continuous in y, and differentiable almost surely in θ_0.[1] Then, we can approximate $\frac{\partial Y^*}{\partial \theta_0}$ via

$$-\frac{\sum_{i=1}^{N} \frac{\partial G(y;\mathbb{Z}(\theta_0)^i,\theta_0)}{\partial \theta_0}}{\sum_{i=1}^{N} g(y;\mathbb{Z}(\theta_0)^i,\theta_0)}\Bigg|_{y=Y^*} \tag{2}$$

where $\mathbb{Z}(\theta_0)^i \sim f_{\mathbb{Z}}(\cdot;\theta_0), i = 1,\ldots,N$ and g is the derivative of G with respect to y.

[1] Note that we make the dependence of $\mathbb{Z}$ on θ_0 explicit.

The Model

All scenarios are based within the linear regressions framework

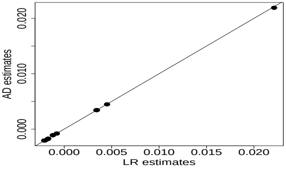
$$y_i = x_i' \boldsymbol{\beta} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma_i^2).$$

we fix an independent conjugate Normal prior for $\boldsymbol{\beta} \sim N_k(\mathbf{b}_0, \mathbf{B}_0)$, consider different error distributions and prior scenarios for σ_i^2 that give rise to different MCMC sampling schemes

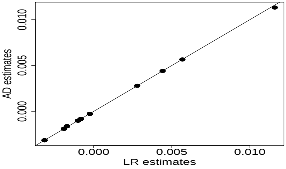
Table: Various set-ups considered for comparative sensitivity analysis via MCMC AD approach and LR method.

Model	Prior (σ^2)	K	p	Sampler
$\sigma_i^2 = \sigma^2$	5	22	$\sigma^{-2} \sim G\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right)$	Gibbs (N-G)
	$\sigma^2 \sim LN(\mu_0, \zeta_0)$	5	22	Gibbs (N), Slice
	$\sigma^2 \sim LN(\mu_0, \zeta_0)$	5	22	Gibbs (N), MH
$\epsilon_i \sim N(0, \lambda_i^{-1}\sigma^2), \lambda_i \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$	$\sigma^{-2} \sim G\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right)$	305	23	Gibbs (N-G-G)

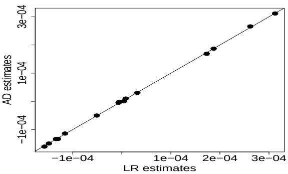
Convergence of Sensitivity Estimates



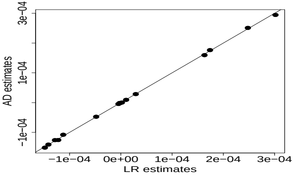
(a) Normal: Gibbs



(b) Student-t: Gibbs



(c) LN : Gibbs and Slice



(d) LN: Gibbs and MH sampler

Stability of Sensitivity Estimates

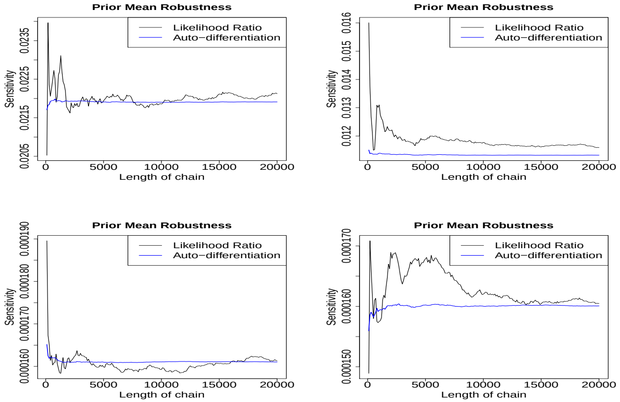


Figure: Convergence and Stability of AD and LR Sensitivity Estimates for $\partial\hat{\beta}_1$ and $\partial\hat{\beta}_1$ for $\dim(\beta) = 4$

Vector Autoregressive Models

A vector autoregression (VAR) is a multiple-equation linear regression that aims to capture the linear interdependencies between variables over time. More specifically, let $\mathbf{y}_t$ denote a vector of observations of n variables at time t with $t = 1, \ldots, T$. Then, a p-order VAR, denoted as VAR(p), is given by:

$$\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{B}_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad (3)$$

where $\mathbf{b}$ is an $n \times 1$ vector of intercepts, $\mathbf{B}_1, \ldots, \mathbf{B}_p$ are $n \times n$ matrices of VAR coefficients and $\boldsymbol{\Sigma}$ is a covariance matrix.

Shrinkage via Minnesota Prior

Minnesota-type prior that shrinks the VAR coefficients to zero. Specifically, we set $\boldsymbol{\beta}_0 = \mathbf{0}$, and the covariance matrix $\mathbb{V}_{\boldsymbol{\beta}}$ is assumed to be diagonal with diagonal elements $v_{\boldsymbol{\beta},ii} = \kappa_1/(l^2 \hat{s}_r)$ for a coefficient associated to lag l of variable r and $v_{\boldsymbol{\beta},ii} = \kappa_2$ for an intercept, where $\hat{s}_r$ is the sample variance of an AR(4) model for the variable r. Further we set $\nu_0 = n + 3$, $\mathbb{S}_0 = \kappa_3 \mathbf{I}_n$, $\kappa_1 = 0.2^2$, $\kappa_2 = 10^2$ and $\kappa_3 = 1$.

Forecast

Even though neither the predictive mean nor any predictive quantiles are available analytically, they can be easily estimated using simulation. Note that the predictive distribution at time $t+h$ can be expressed as

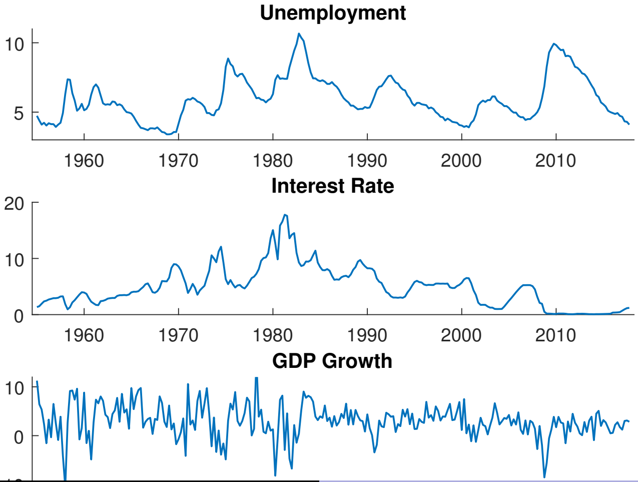
$$p(\mathbf{y}_{t+h}|\mathbf{y}_{1:t}) = \int p(\mathbf{y}_{t+h}|\mathbf{y}_{1:t}, \boldsymbol{\beta}, \boldsymbol{\Sigma})p(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}_{1:t})\mathrm{d}(\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where $p(\mathbf{y}_{t+h}|\mathbf{y}_{1:t}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ is a Gaussian density implied by the Gaussian VAR.

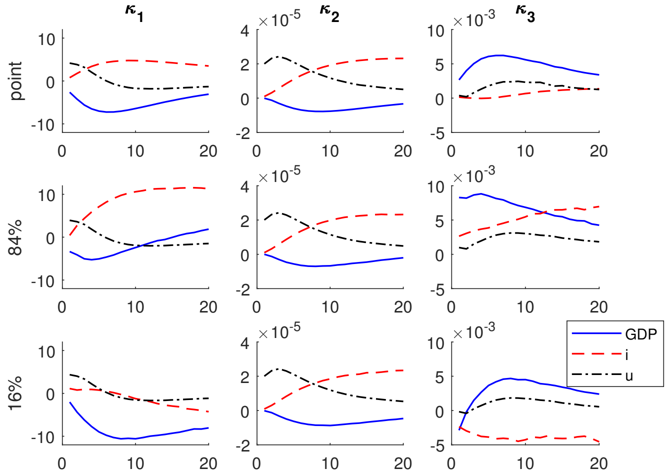
Generate $\mathbf{y}^g_{t+h}$ from

$$(\mathbf{y}^g_{t+h}|\mathbf{y}_{1:t}, \boldsymbol{\beta}^g, \boldsymbol{\Sigma}^g) \sim \mathcal{N}(\mathbf{X}_{t+h}\boldsymbol{\beta}^g, \boldsymbol{\Sigma}^g).$$

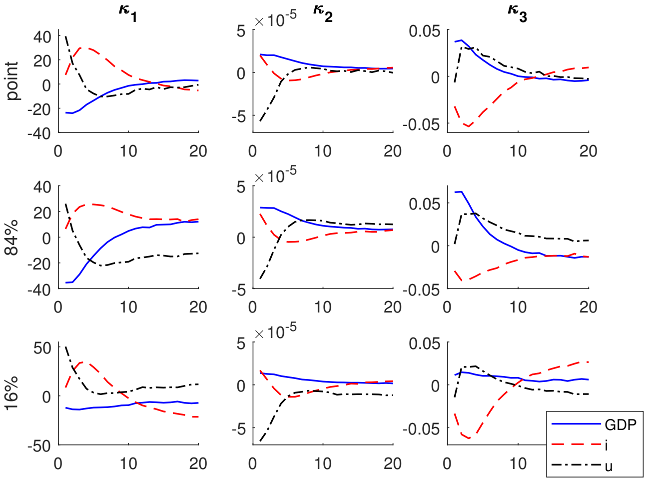
US quarterly data from 1954:Q3 to 2017:Q4



Sensitivities for the Minnesota Prior



Sensitivities for Sub-sample



The gold standard for Bayesian model comparison is the Bayes factor. Specifically, the *Bayes factor* in favor of M_i against M_j is defined as

$$\mathrm{BF}_{ij} = \frac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)},$$

where

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|\psi_k, M_k)p(\psi_k|M_k)d\psi_k \tag{4}$$

is the *marginal likelihood* under model M_k, $k = i, j$.

Model Comparison

To see the effect of perturbation analysis via AD, we have

$$BF_{i,j}(\eta_0^{'}, \eta_i^{'}, \eta_j^{'}) \approx BF_{i,j}(\eta_0, \eta_i, \eta_j) + \nabla BF_{i,j}(\eta_0, \eta_i, \eta_j)^T \begin{bmatrix} \eta_0' - \eta_0 \\ \eta_i' - \eta_i \\ \eta_i' - \eta_i \end{bmatrix} \tag{5}$$

where three partial derivative vectors are computed simultanenous via AD irregardless of the perturbation size.

Table: Log marginal likelihood estimates of the VAR and VAR with t innovations using the cross-entropy method (CE) and Chib's method (Chib).

	VAR	VAR-t		
		$\nu = 5$	$\nu = 10$	$\nu = 30$
CE	-1416.7	-1322.2	-1344.7	-1381.5
Chib	-1416.7	-1322.2	-1344.7	-1381.5

Table: Derivatives of log marginal likelihood estimates of the VAR and VAR with t innovations with respect to the hyperparameters.

	VAR			VAR-t ($\nu = 5$)		
	κ_1	κ_2	κ_3	κ_1	κ_2	κ_3
CE	424.3	-0.01	10.3	471.7	-0.01	5.6
Chib	424.3	-0.01	10.3	471.8	-0.01	5.6

Optimal Prior

We sometimes also interested in the "optimal" prior, i.e.

$$\boldsymbol{\eta}_0^* = \arg\max_{\boldsymbol{\eta}_0} p(\mathbf{y}; \boldsymbol{\eta}_0).$$

Derivatives of the marginal likelihood can greatly enhance the optimization procedure.

Large Bayesian VARs with the natural conjugate prior are now routinely used for forecasting and structural analysis. More specifically, the marginal distribution on $\mathbf{\Sigma}$ is inverse-Wishart and the conditional distribution on $\mathbb{A}$ is normal:

$$\mathbf{\Sigma} \sim \mathcal{IW}(\nu_0, \mathbb{S}_0), \quad (\text{vec}(\mathbb{A})|\mathbf{\Sigma}) \sim \mathcal{N}(\text{vec}(\mathbb{A}_0), \mathbf{\Sigma} \otimes \mathbb{V}_{\mathbb{A}}),$$

and we write $(\mathbb{A}, \mathbf{\Sigma}) \sim \mathcal{NIW}(\mathbb{A}_0, \mathbb{V}_{\mathbb{A}}, \nu_0, \mathbb{S}_0)$.

We set $\mathrm{vec}(\mathbb{A}_0) = \mathbf{0}$ to shrink the VAR coefficients to zero, and $\mathbb{V}_{\mathbb{A}}$ to be diagonal with the i-th diagonal element $v_{\mathbb{A},ii}$ set as:

$$v_{\mathbb{A},ii} = \begin{cases} \frac{\kappa_1}{l^{\kappa_2} s_r^2}, & \text{for the coefficient on the } l\text{-th lag of variable } r \\ \kappa_3, & \text{for an intercept} \end{cases}$$

where s_r^2 is the sample variance of the residuals from an AR(p) model for the variable r. Hence, we simplify the task of eliciting $\mathbb{V}_{\mathbb{A}}$ by choosing only three key hyperparameters κ_1, κ_2 and κ_3. For $\mathbf{\Sigma}$, we introduce two additional hyperparameters κ_4 and κ_5, and set $k_{0,\mathbf{\Sigma}} = \kappa_4 + n + 1$ and $\mathbb{S}_{0,\mathbf{\Sigma}} = \kappa_5 \mathrm{diag}(s_1^2, \ldots, s_n^2)$.

Table: Baseline and optimized values of the hyperparameters.

	baseline	optimize κ_1-κ_3	optimize κ_1-κ_5
κ_1	0.05	0.051	0.041
κ_2	1	3.2	3.2
κ_3	100	28.2	24.2
κ_4	1	1	13.0
κ_5	1	1	10.3
log-ML	11,093	11,216	11,395

The dataset for our forecasting exercise consists of 18 US quarterly variables and covers the quarters from 1959Q1 to 2018Q4. It is sourced from the FRED-QD database at the Federal Reserve Bank of St.

Table: Computation times of the proposed AD approach and the grid-search approach (in seconds). The numbers in parenthesis are the numbers of grid points in each dimension for the grid-search approach.

optimize κ_1-κ_3			optimize κ_1-κ_5		
grid (30)	grid (60)	AD	grid (30)	grid (60)	AD
17.8	138	27.2	16,020	496,800	32.0

Conclusion

We applied the AD approach to provide comprehensive assessment of sensitivity in Bayesian MCMC analysis, effective

- provide guidelines for choices of prior parameters
- give clear indication on the convergence

The results obtained via AD agrees with the existing LR method, but at a faster convergence as well as efficiency after convergence.
Future directions:

- Use the sensitivity approach to assess efficiency of the sampler after convergence.
- Extend the method to cases where the full conditional distribution is unknown, i.e. MH and sequential Monte-Carlo.