Convergence Complexity of Gibbs Samplers for Bayesian Vector Autogressions

Galin Jones[1]

University of Minnesota
galin@umn.edu
@JonesGalin

SDSS 2020

[1] Joint work with Karl Oskar Ekvall @KarlOskar

Bayesian VARX

r-valued stochastic process

$$Y_t = \sum_{i=1}^{q} \mathcal{A}_i^T Y_{t-i} + \mathcal{B}^T X_t + \varepsilon_t \qquad \varepsilon_t \overset{ind}{\sim} N_r(0, \Sigma)$$

$\{X_t\}$ indep. $\{\varepsilon_t\}$ and distribution not depending on $\{\mathcal{A}_i\}, \mathcal{B}, \Sigma$

Bayesian VARX

r-valued stochastic process

$$Y_t = \sum_{i=1}^{q} \mathcal{A}_i^T Y_{t-i} + \mathcal{B}^T X_t + \varepsilon_t \qquad \varepsilon_t \overset{ind}{\sim} N_r(0, \Sigma)$$

$\{X_t\}$ indep. $\{\varepsilon_t\}$ and distribution not depending on $\{\mathcal{A}_i\}, \mathcal{B}, \Sigma$

Rewrite

$$Y_t = \mathcal{A}^T Z_t + \mathcal{B}^T X_t + \varepsilon_t$$

$\mathcal{A} = [\mathcal{A}_1^T, \ldots, \mathcal{A}_q^T] \in \mathbb{R}^{qr \times r}, \quad \mathcal{B} \in \mathbb{R}^{p \times r}$

$Z_t = [Y_{t-1}^T, \ldots, Y_{t-q}^T]^T \in \mathbb{R}^{qr}, \quad Z_1$ is fixed

Bayesian VARX

r-valued stochastic process

$$Y_t = \sum_{i=1}^{q} \mathcal{A}_i^T Y_{t-i} + \mathcal{B}^T X_t + \varepsilon_t \qquad \varepsilon_t \overset{ind}{\sim} N_r(0, \Sigma)$$

$\{X_t\}$ indep. $\{\varepsilon_t\}$ and distribution not depending on $\{\mathcal{A}_i\}, \mathcal{B}, \Sigma$

Rewrite

$$Y_t = \mathcal{A}^T Z_t + \mathcal{B}^T X_t + \varepsilon_t$$

$\mathcal{A} = [\mathcal{A}_1^T, \ldots, \mathcal{A}_q^T] \in \mathbb{R}^{qr \times r}, \quad \mathcal{B} \in \mathbb{R}^{p \times r}$

$Z_t = [Y_{t-1}^T, \ldots, Y_{t-q}^T]^T \in \mathbb{R}^{qr}, \quad Z_1$ is fixed

small-n: Data fixed, $n > p$ but possibly small compared to qr

large-n: Data stochastic, $n > p$, and n is increasing

Likelihood

$$Y_t = \mathcal{A}^T Z_t + \mathcal{B}^T X_t + \varepsilon_t \qquad \varepsilon_t \overset{ind}{\sim} N_r(0, \Sigma)$$

$$S = n^{-1}(Y_t - \mathcal{A}^T Z_t - \mathcal{B}^T X_t)^T (Y_t - \mathcal{A}^T Z_t - \mathcal{B}^T X_t)$$

$$f(Y, X | \mathcal{A}, \mathcal{B}, \Sigma) \propto |\Sigma|^{-n/2} \text{etr}\left(-\frac{n}{2} S \Sigma^{-1}\right)$$

Instead of $\mathcal{A}$ it is common to work with $\alpha = \text{vec}(\mathcal{A}) \in \mathbb{R}^{qr^2}$. Thus we need a prior for

$$(\alpha, \mathcal{B}, \Sigma) \in \mathbb{R}^{qr^2} \times \mathbb{R}^{p \times r} \times S_{++}^r$$

VARX Priors

$$Y_t = \mathcal{A}^T Z_t + \mathcal{B}^T X_t + \varepsilon_t \qquad \varepsilon_t \overset{ind}{\sim} N_r(0, \Sigma)$$

Karlsson (2013, Handbook Economic Forecasting) gives a comprehensive review of priors.

Let $\Psi = [\mathcal{A}^T, \mathcal{B}^T]^T$.

vec(Ψ) $\sim$ MVN and $\Sigma \sim$ Inverse Wishart,

$f(\Psi, \Sigma) \propto |\Sigma|^{-a}$,

Minnesota prior, and so on.

Proposed prior: Recall $\alpha = \text{vec}(\mathcal{A}) \in \mathbb{R}^{qr^2}$

$$f(\alpha) \qquad f(\mathcal{B}) \propto 1 \qquad f(\Sigma) \propto |\Sigma|^{-a/2} \text{etr}\left(-\frac{1}{2} D \Sigma^{-1}\right)$$

VARX Priors

$\alpha = \text{vec}(\mathcal{A}) \in \mathbb{R}^{qr^2}$ and

$$f(\alpha) \qquad f(\mathcal{B}) \propto 1 \qquad f(\Sigma) \propto |\Sigma|^{-a/2}\text{etr}\left(-\frac{1}{2}D\Sigma^{-1}\right)$$

Thm If either

1. D is positive definite, X has full column rank, $n + a > 2r + p$, and $f(\alpha)$ is proper; or
2. $[Y, Z, X]$ has full column rank, $n + a > (2 + q)r + p$, and $f(\alpha)$ is bounded,

then the posterior $f(\alpha, \mathcal{B}, \Sigma | \mathcal{D}_n)$ exists and is proper.

VARX Posterior

$$f(\mathcal{B}) \propto 1 \qquad f(\Sigma) \propto |\Sigma|^{-a/2} \mathrm{etr}\left(-\frac{1}{2} D \Sigma^{-1}\right)$$

$$f(\alpha) \propto \exp\left(-\frac{1}{2}(\alpha - m)^T C (\alpha - m)\right)$$

$f(\alpha)$ common in macroeconomics and finance

allows large VARs, i.e. qr large and m can be chosen to address near non-stationarity (unit root sense)

$f(\mathcal{B})$ common in multivariate location-scale settings

$f(\Sigma)$ includes inverse Wishart and Jeffreys priors

Collapsed Gibbs sampler

(α, Σ) is a linchpin variable:

$$f(\alpha, \mathcal{B}, \Sigma | \mathcal{D}_n) = f(\mathcal{B} | \alpha, \Sigma, \mathcal{D}_n) f(\alpha, \Sigma | \mathcal{D}_n)$$

and

$$\mathcal{B} \mid \alpha, \Sigma, \mathcal{D}_n \sim \text{Matrix Normal}$$

Use Gibbs sampler for $f(\alpha, \Sigma | \mathcal{D}_n)$ since

$$\Sigma \mid \alpha, \mathcal{D}_n \sim \text{Inverse Wishart}$$
$$\alpha \mid \Sigma, \mathcal{D}_n \sim \text{Multivariate Normal}$$

$$\theta = (\mathcal{B}, \Sigma, \alpha) \to (\mathcal{B}, \Sigma', \alpha) \to (\mathcal{B}, \Sigma', \alpha') \to (\mathcal{B}', \Sigma', \alpha') = \theta'$$

Convergence Analysis

Geometric ergodicity of Collapsed Gibbs sampler: Find $\rho < 1$ s.t.

$$\|K_C^h(\theta, \cdot) - F(\cdot|\mathcal{D}_n)\|_{TV} \leq M(\theta)\rho^h$$

Classical (small n):
Find conditions to ensure that geometric ergodicity holds for any fixed data set.

Convergence Analysis

Geometric ergodicity of Collapsed Gibbs sampler: Find $\rho < 1$ s.t.

$$\|K_C^h(\theta, \cdot) - F(\cdot|\mathcal{D}_n)\|_{TV} \leq M(\theta)\rho^h$$

Classical (small n):
Find conditions to ensure that geometric ergodicity holds for any fixed data set.

Convergence complexity (large n): $\rho = \rho_n$
Find conditions to ensure that the convergence rate behaves well for large n:

$$\limsup_{n\to\infty} \rho_n < 1 \quad \text{almost surely}$$

so the geometric ergodicity is *asymptotically stable*.

Convergence Analysis

Collapsed Gibbs sampler:

$$\theta = (\mathcal{B}, \Sigma, \alpha) \to (\mathcal{B}, \Sigma', \alpha) \to (\mathcal{B}, \Sigma', \alpha') \to (\mathcal{B}', \Sigma', \alpha') = \theta'$$

It suffices to study the marginal process $\{\alpha^h\}$ because

$$\|K_C^h(\theta, \cdot) - F(\cdot|\mathcal{D}_n)\|_{TV} \leq \|K_\alpha^{h-1}(\alpha, \cdot) - F_\alpha(\cdot|\mathcal{D}_n)\|_{TV}$$

Convergence Analysis

Rosenthal (JASA, 1995)

Suppose $V : \mathbb{R}^{qr^2} \to [0, \infty)$, $\lambda < 1$ and some $L < \infty$

$$\int V(\alpha) K_\alpha(\alpha', d\alpha) \leq \lambda V(\alpha') + L \qquad \text{for all } \alpha'. \tag{1}$$

and for $T > 2L/(1 - \lambda)$

$$K_\alpha(\alpha, \cdot) \geq \varepsilon R(\cdot) \qquad \text{for all } \alpha \in \{\alpha : V(\alpha) \leq T\}. \tag{2}$$

Then K_α is geometrically ergodic and there is an explicit formula for

$$\bar{\rho} = \bar{\rho}_n(Y, X, C, D, m, a)$$

such that

$$\rho \leq \bar{\rho}$$

Convergence Analysis

Take home message:

Rosenthal's theorem yields an explicit upper bound on the rate

$$\bar{\rho} = \bar{\rho}(\mathcal{D}_n, C, D, m, a),$$

but if V is not chosen carefully, then it is likely the case that

$$\liminf_{n \to \infty} \bar{\rho} \to 1 \quad \text{almost surely}$$

and we won't be able to conclude that the geometric ergodicity is asymptotically stable.

Classical (small n) Convergence Analysis

Recall

$$f(\alpha) \propto \exp\left(-\frac{1}{2}(\alpha - m)^T C(\alpha - m)\right)$$

<u>Thm</u> If C is positive definite, then the collapsed Gibbs sampler is geometrically ergodic.

Classical (small n) Convergence Analysis

Recall

$$f(\alpha) \propto \exp\left(-\frac{1}{2}(\alpha - m)^T C(\alpha - m)\right)$$

Thm If C is positive definite, then the collapsed Gibbs sampler is geometrically ergodic.

But

$$\liminf_{n\to\infty} \bar{\rho}_n = 1 \quad \text{almost surely}$$

Classical (small n) Convergence Analysis

Recall

$$f(\alpha) \propto \exp\left(-\frac{1}{2}(\alpha - m)^T C(\alpha - m)\right)$$

Thm If C is positive definite, then the collapsed Gibbs sampler is geometrically ergodic.

But

$$\liminf_{n\to\infty} \bar{\rho}_n = 1 \quad \text{almost surely}$$

Used drift function

$$V(\alpha) = \|\alpha\|^2$$

which implies that the Markov chain should visit sets near the origin frequently.

Classical (small n) Convergence Analysis

Recall

$$f(\alpha) \propto \exp\left(-\frac{1}{2}(\alpha - m)^T C(\alpha - m)\right)$$

Thm If C is positive definite, then the collapsed Gibbs sampler is geometrically ergodic.

But

$$\liminf_{n\to\infty} \bar{\rho}_n = 1 \quad \text{almost surely}$$

Used drift function

$$V(\alpha) = \|\alpha\|^2$$

which implies that the Markov chain should visit sets near the origin frequently.

No reason to think this is reasonable when n is large.

Convergence Complexity (large *n*) Analysis

Intuition:
For large *n*, the posterior should concentrate around the true α^*.

Convergence Complexity (large n) Analysis

Intuition:
For large n, the posterior should concentrate around the true α^*.

For large n, the least squares estimator or MLE

$$\hat{\mathcal{A}} = (Z^T Q_X Z)^+ Z^T Q_X Y$$

should converge to the true α^*.

Convergence Complexity (large n) Analysis

Intuition:
For large n, the posterior should concentrate around the true α^*.

For large n, the least squares estimator or MLE

$$\hat{\mathcal{A}} = (Z^T Q_X Z)^+ Z^T Q_X Y$$

should converge to the true α^*.

Maybe we should center the drift function around the least squares estimator or MLE.

$$\begin{aligned} V(\alpha) &= \|Q_X Z \mathcal{A} - Q_X Z \hat{\mathcal{A}}\|_F^2 \\ &= \|(I_r \otimes Q_X Z)(\alpha - \hat{\alpha})\|^2 \end{aligned}$$

Convergence Complexity (large n) Analysis

Thm If

(a) C is positive definite,

(b) $[Y, X, Z]$ has full column rank for all large enough n almost surely,

(c) $\|\hat{\alpha}\|^2 = O(1)$ almost surely as $n \to \infty$, and

(d) there exists $0 \leq M < \infty$ such that, almost surely,

$$\begin{aligned} M^{-1} &\leq \liminf_{n\to\infty} n^{-1}\lambda_{min}(Y^T Q_{[Z,X]} Y) \\ &\leq \limsup_{n\to\infty} n^{-1}\lambda_{max}(Y^T Q_{[Z,X]} Y) \\ &\leq M \end{aligned}$$

then, almost surely,

$$\limsup_{n\to\infty} \bar{\rho}_n < 1$$