An elliptical slice sampler for regression models with shrinkage priors

Jingyu He
with P. Richard Hahn
SBIES 2016

April 30, 2016

Outline

- An elliptical slice sampler for Bayesian regression with general shrinkage priors.
- Simulation of horseshoe regression.
- Application to IV models.

Elliptical Slice Sampler for Regression

- The original elliptical slice sampler (Murray et. al. [2010]) is designed to sampling from posterior with normal prior and general likelihood.
- It's also applicable to normal likelihood and general prior such as shrinkage priors.
- Advantages:
  - It only requires evaluating the prior density (or an approximation.)
  - Sample all coefficients simultaneously. Not necessary to loop over variables.

Elliptical Slice Sampler for Regression

- Regression model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$$

$$\epsilon \sim N(0, \sigma^2), \beta \sim \pi(\beta)$$

- Posterior

$$\underbrace{f(\boldsymbol{y} \mid \beta, \sigma^2)}_{\text{normal}} \quad \underbrace{\pi(\beta)}_{\text{arbitrary prior}}$$

- Sampling β from normal distribution $p_f(\beta \mid \mathbf{y}, \mathbf{x}, \sigma^2) \propto f(\boldsymbol{y} \mid \beta, \sigma^2)$, a posterior of β under flat prior.

- Update σ^2 by conjugate inverse-Gamma in between sampling β.

Elliptical Slice Sampler for Regression

Elliptical slice sampler for shrinkage regression

1. Start from a initial value β.
2. Choose ellipse $\boldsymbol{\nu} \sim N(0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$
3. Set log-prior threshold, $u \sim \text{Uniform}[0,1]$, $\log y \leftarrow \log \pi(\beta) + \log u$
4. Draw an initial proposal sample, $\theta \sim \text{Uniform}[0, 2\pi]$, $[\theta_{\text{min}}, \theta_{\text{max}}] \leftarrow [\theta - 2\pi, \theta]$
5. Let $\beta' \leftarrow \beta\cos(\theta) + \boldsymbol{\nu}\sin(\theta)$
6. If $\log \pi(\beta') > \log y$, then accept, return β'
7. else, shrink the bracket and sample another point:
   - 7.1 if $\theta < 0$, then $\theta_{\text{min}} \leftarrow \theta$, else $\theta_{\text{max}} \leftarrow \theta$
   - 7.2 Draw proposal again : $\theta \sim \text{Uniform}[\theta_{\text{min}}, \theta_{\text{max}}]$
   - 7.3 Go to 4.

Example

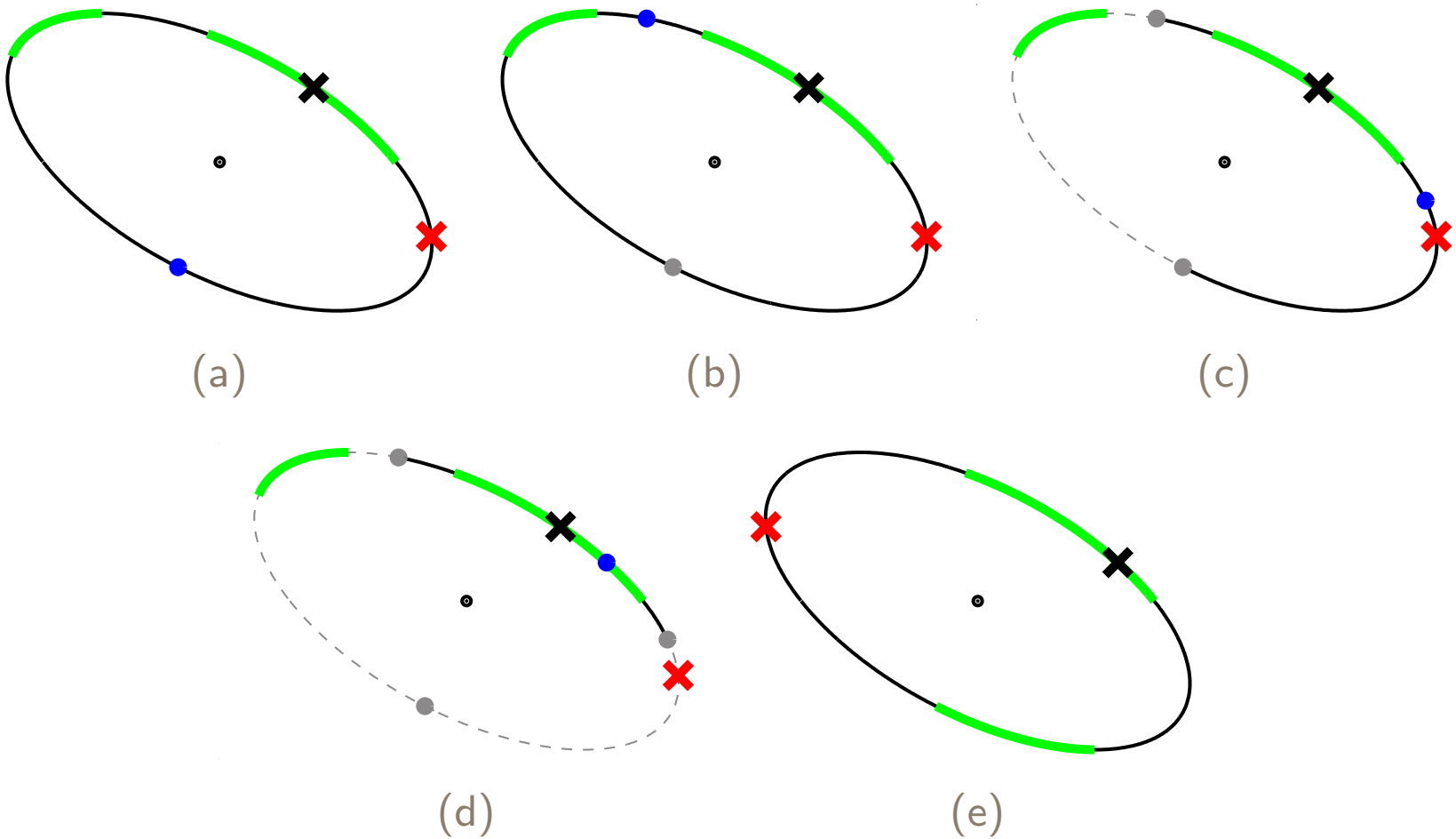

Figure: Illustration of the elliptical slice sampler

(Approximate) Horseshoe prior

The popular "horseshoe" prior [Carvalho (2010)] is a local scale-mixture of normals

$$\boldsymbol{\delta} \sim \mathrm{N}(0, \lambda_0^2 \boldsymbol{\Lambda}^2),\ \lambda_j \sim \mathrm{C}^+(0,1),\ \lambda_0 \sim \mathrm{C}^+(0,1).$$

The horseshoe prior lacks a closed-form density, but has a good approximation

$$\frac{K}{2} \log\left(1 + \frac{4}{(\delta_j/\lambda_0)^2}\right) < \pi(\delta_j/\lambda_0) < K \log\left(1 + \frac{2}{(\delta_j/\lambda_0)^2}\right) \qquad (1)$$

where $K = 1/(2\pi^3)^{1/2}$.

We use the easy-to-evaluate density $\pi(\delta_j/\lambda_0) \propto \log\left(1 + \frac{4}{(\delta_j/\lambda_0)^2}\right)$.

Simulations

We compare the elliptical slice sampler with two Gibbs samplers

- R package `monomvn`, standard Gibbs sampler. It was written in C++ by Robert B. Gramacy.
- Makalic et. al (2016) propose another Gibbs sampler with auxiliary variables. Makalic provides Matlab code, we rewrote it to C++.

Simulations

Data is generated by the horseshoe prior. $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are length p vectors. $\mathbf{X}$ is a $n \times p$ matrix. C^+ stands for half-Cauchy distribution.

$$
\begin{aligned}
\sigma &= 1 \\
\tau &\sim C^+(0, 1) \\
\lambda_i &\sim C^+(0, 1) \\
\beta_i &\sim N(0, \tau \times \lambda_i \times \sigma) \\
\mathbf{X} &\sim N(0, 1) \\
\mathbf{y} &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma)
\end{aligned}
\tag{2}
$$

We care about

- Whether the elliptical slice sampler get similar estimation as Gibbs sampler.
- Computing time.

Simulations, MSE

n is number of observations, p is number of variables. 13,000 posterior samples are drawn and 3,000 of them are burn-in samples.

n	p	MSE			
		Makalic		Slice	`monomvn`
		C++	Matlab		
1000	20	0.020	0.019	0.020	0.020
5000	20	0.004	0.004	0.004	0.004
1000	50	0.049	0.052	0.049	0.049
5000	50	0.010	0.010	0.010	0.010
1000	100	0.105	0.108	0.107	0.105
5000	100	0.020	0.020	0.020	0.020
1000	500	0.975	0.970	0.986	0.973
5000	500	0.110	0.112	0.111	0.110
5000	1000	0.249	0.249	0.249	0.249

Table: MSE, drawing 13,000 posterior samples

Simulations, Computing time

n is number of observations, p is number of variables. 13,000 posterior samples are drawn and 3,000 of them are burn-in samples.

n	p	Running time (in seconds)			
		Makalic		Slice	`monomvn`
		C++	Matlab		
1000	20	1.56	4.85	0.25	1.80
5000	20	4.53	7.05	0.23	7.64
1000	50	3.14	5.33	0.41	2.36
5000	50	7.52	13.76	0.39	6.81
1000	100	12.20	9.04	1.49	6.75
5000	100	23.33	28.08	1.26	12.28
1000	500	262.82	96.51	8.58	176.50
5000	500	330.79	162.69	8.51	213.63
5000	1000	1478.59	483.73	32.76	1214.59

Table: Running time, drawing 13,000 posterior samples. 3,000 burn-in samples

Simulations, Effective Sample Size

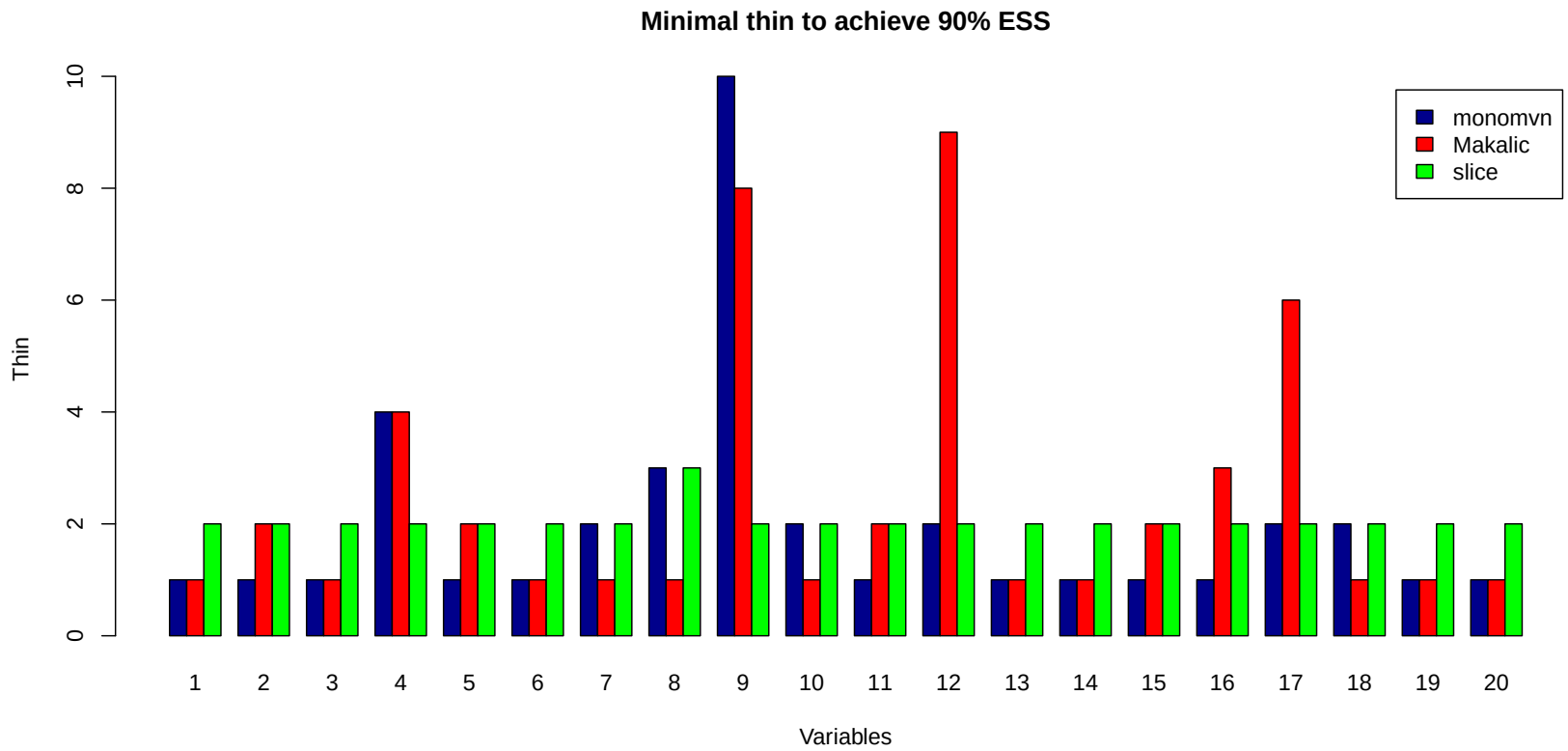

Figure: Minimal among of thinning to achieve 90% ESS for each variable, good data case.

Simulations, Effective Sample Size

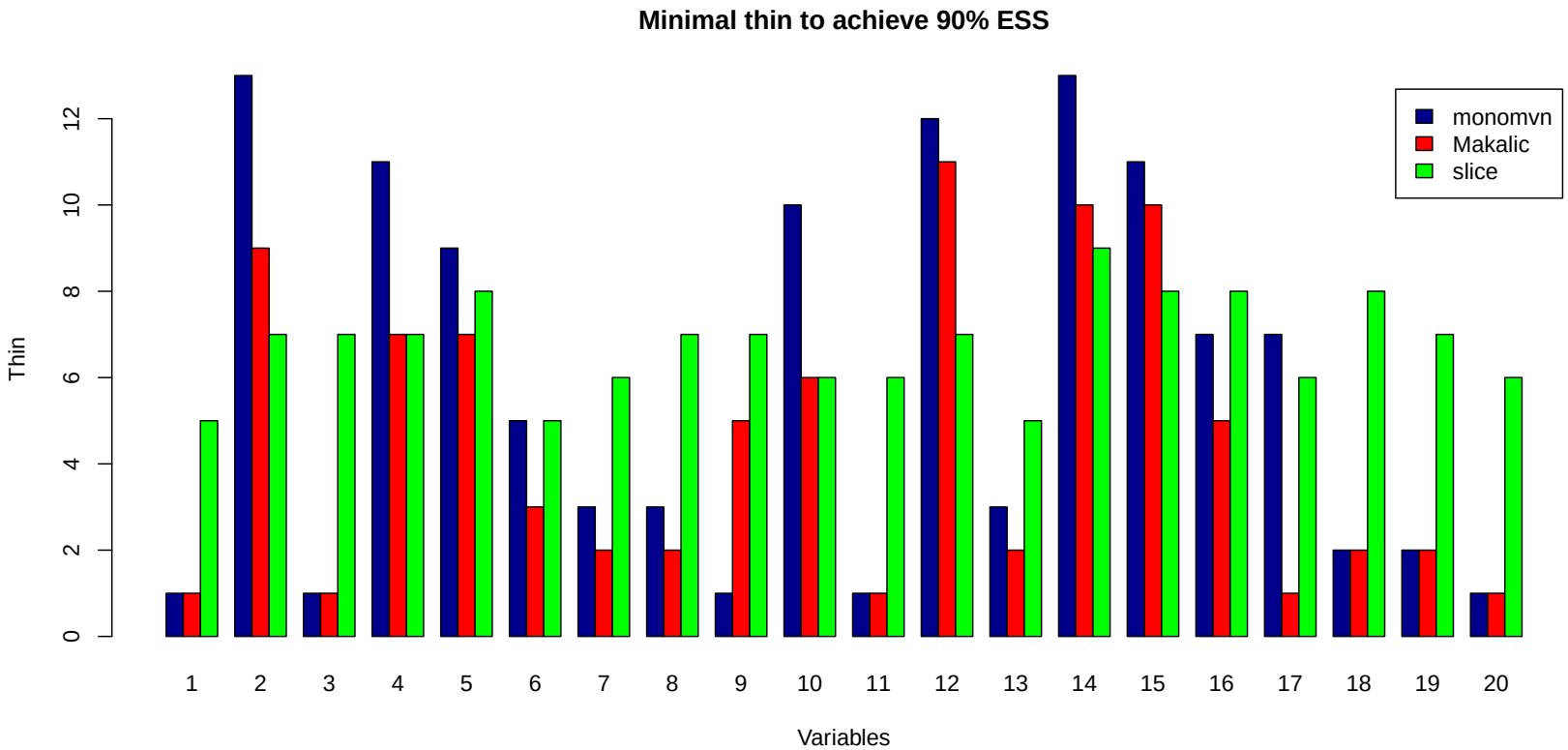

Figure: Minimal among of thinning to achieve 90% ESS for each variable, bad data case

Summary

Our method is an order of magnitude faster per sample and requires less thinning (fewer total samples) than the standard Gibbs samplers.

Application to IV Model

Here is a way to write IV models

$$\begin{aligned} \mathbf{y} &= \mathbf{x}(\alpha + \beta) - \alpha \mathbf{Z}\delta + \xi\varepsilon \\ \mathbf{x} &= \mathbf{Z}\delta + \sigma\epsilon \end{aligned} \tag{3}$$

We reparametrize as $\beta^* = (\alpha + \beta)$, $\delta^* = \alpha\delta$, and $\alpha^* = 1/\alpha$

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\beta^* - \mathbf{Z}\delta^* + \xi\varepsilon \\ \mathbf{x} &= \mathbf{Z}\delta^*\alpha^* + \sigma\epsilon \end{aligned} \tag{4}$$

δ^* appears in both equations, exclusive restriction.

Gibbs update for β^* and δ^*

To deploy the slice sampler, we write the likelihood as

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\beta^* - \mathbf{Z}\delta^* + \xi\varepsilon \\ \frac{\xi\mathbf{x}}{\sigma} &= \mathbf{Z}\delta^*\alpha^* \cdot \frac{\xi}{\sigma} + \xi\epsilon \end{aligned} \tag{5}$$

Therefore, let

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{x} \\ \mathbf{y} \end{pmatrix}, \tilde{\mathbf{Z}} = \begin{pmatrix} \frac{\xi}{\sigma}\alpha^*\mathbf{Z} & 0 \\ -\mathbf{Z} & \mathbf{x}, \end{pmatrix}, \tilde{\delta} = \begin{pmatrix} \delta^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} \alpha\delta \\ \alpha + \beta \end{pmatrix} \tag{6}$$

and configure the regression as

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{Z}}\tilde{\delta} + \xi\varepsilon \tag{7}$$

- We sample $\tilde{\delta}$ by elliptical slice sampler, then update ξ, σ and $\alpha*$ by regression with inverse-gamma and a normal prior respectively.
- Priors are placed directly on the new parameter space.

Returns to schooling data

Angrist and Krueger investigate the impact of schooling on wages, the data is from the 1980 U.S. Census on men born between 1930 and 1939.

We follow the analysis of Hansen and Kozbur (2014). 509 control variables, consisting of 9 year-of-birth indicators, 50 state-of-birth indicators, as well as the 450 interactions between them.

For instruments, three instruments sets are used

1. 3 quarter-of-birth indicators.
2. Interactions of (1) with the 9 main effects for year-of-birth and 50 main effects for state-of-birth, for a total of 180 instruments.
3. Interactions of (1) with the full set of state-of-birth and year-of-birth controls to obtain a total of 1527 candidate instruments.

Returns to schooling results

	2SLS	Post-LASSO	JIVE	RJIVE	FSP	Slice
	A. 3 instruments					
Coeff.	0.1079	0.115	0.1091	0.1091	0.1098	0.1120
SD	0.0196	0.0205	0.0202	0.0202	0.0207	0.0206
	B. 180 instruments					
Coeff.	0.0928	0.1125	0.1096	0.1062	0.1107	0.1093
SD	0.0097	0.0173	0.0161	0.0157	0.0183	0.0089
	C. 1527 instruments					
Coeff.	0.0712	0.0862	0.0816	0.1067	0.0862	0.0887
SD	0.0049	0.0254	0.5168	0.0171	0.0066	0.0035

Last Slide

Thank you!

Appendix

Empirical Example, Effective Sample Size

`diabetes` data, (Efron et. al. [2004]). 442 observations and 10 regressors.

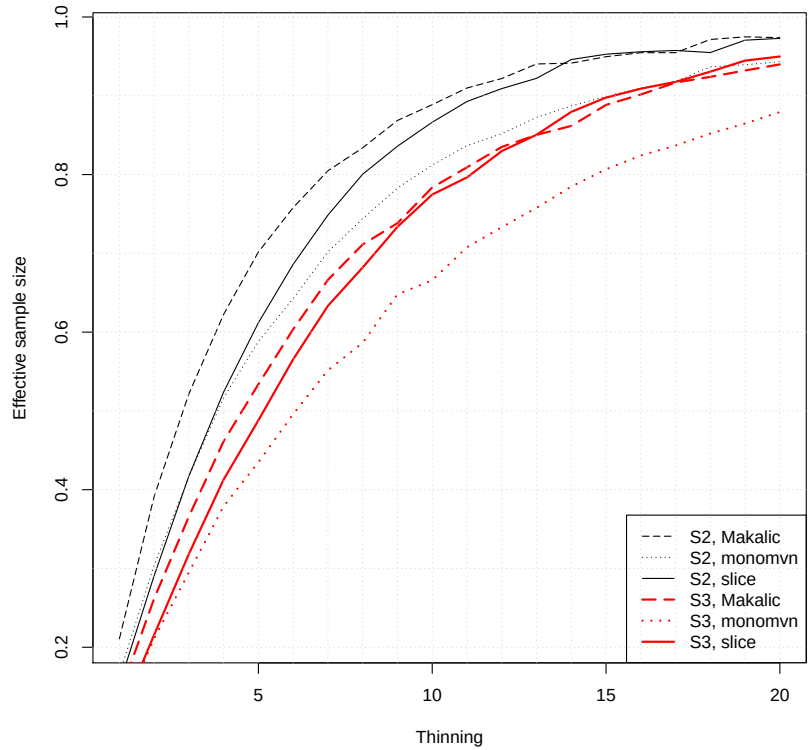

Figure: ESS of Variable `S2` and `S3`, comparing with `monomvn` and Makalic's sampler.

Effective Sample Size

Monte Carlo statistical methods, Robert, Christian and Casella, George, pp 499-500.
Suppose T is sample size. The effective sample size $\hat{T}^S$ is

$$\hat{T}^S = T/\kappa(h)$$

where $\kappa(h)$ is the autocorrelation time associated with the sequence $h(\theta^{(t)})$

$$\kappa(h) = 1 + 2\sum_{t=1}^{\infty} \text{corr}\left(h(\theta^{(0)}), h(\theta^{(t)})\right)$$

Reference

- E. Makalic and D. F. Schmidt, "A Simple Sampler for the Horseshoe Estimator," in *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 179-182, Jan. 2016. doi: 10.1109/LSP.2015.2503725

- Murray, I., R. P. Adams, and D. J. MacKay (2010). Elliptical slice sampling. In *JMLR Workshop and Conference Proceedings*, Volume 9, pp. 541C548. JMLR.

- Gramacy, R.B., Pantaleo, E. (2009). Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. Bayesian Analysis 5(2), pp. 237-262;