Posterior computation for IV regression with shrinkage priors

Jingyu He
Chicago Booth

with P. Richard Hahn (Chicago Booth)
and Hedibert Lopes (INSPER, Brazil)

June 16, 2016

Outline

- An elliptical slice sampler for Bayesian regression with generic shrinkage priors.
- Application to instrumental variables (IV) models.
- The idea of partial factor shrinkage modeling was presented by Hedibert Lopes in EFaB 311 session yesterday.

  More details are available in the paper

  P. Richard Hahn, Jingyu He & Hedibert Lopes (2016): **Bayesian factor model shrinkage for linear IV regression with many instruments**, Journal of Business & Economic Statistics, forthcoming.

Part I : Elliptical Slice Sampler for Regression

Elliptical Slice Sampler for Regression

- The original elliptical slice sampler (Murray et. al. [2010]) was designed to sampling from a posterior arising from a normal prior and a general likelihood.

- It can also be used with a normal likelihood and general prior such as shrinkage priors.

- Advantages:
  - Flexible : It only requires evaluating the prior density or an approximation (no special samplers are required).

  - Fast : 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

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

- Sample β from a 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 a 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.

Visualized illustration, (a)

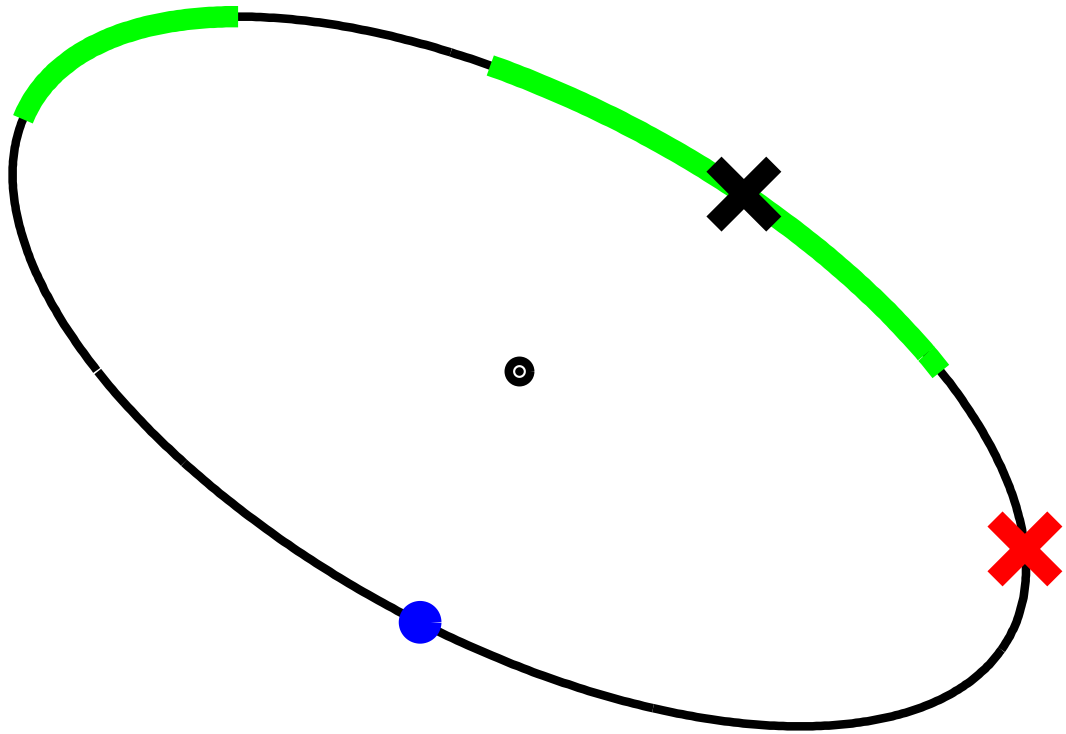

Figure: First round, step 1. ✖ is the initial value. ✖ is the auxiliary value. Green slice contains all samples have larger prior evaluation than the initial value.
The proposal (●) is drawn uniformly on the ellipse by ●$=$✖$\cos(\theta)+$ ✖$\sin(\theta)$

Visualized illustration, (b)

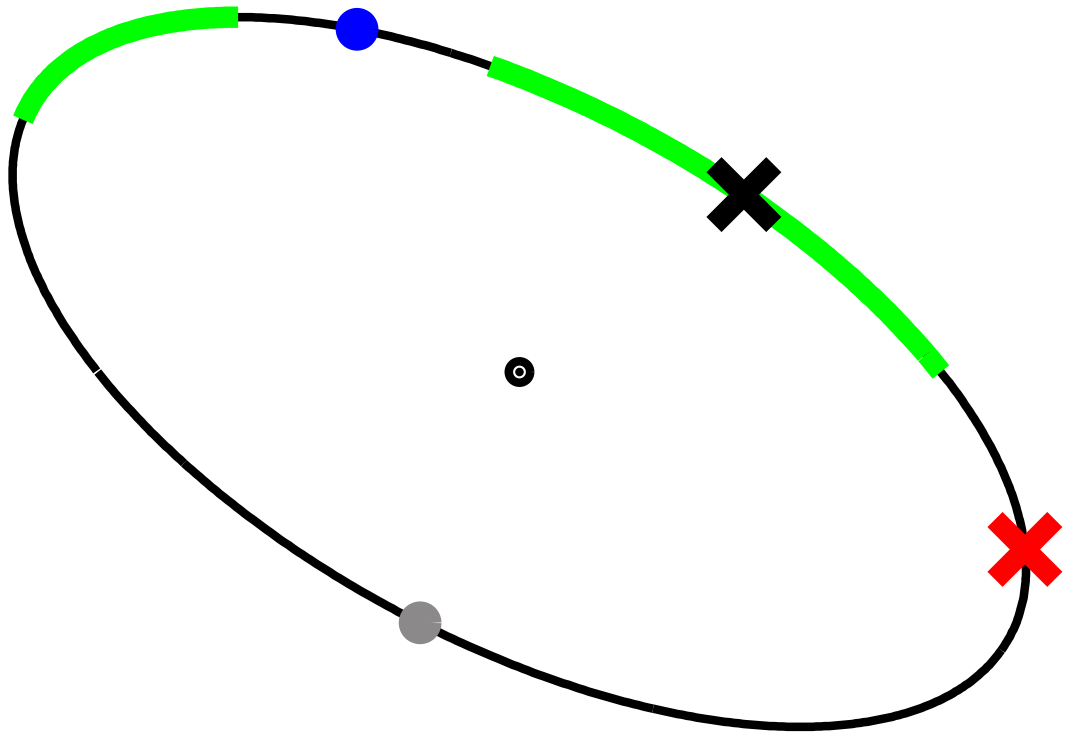

Figure: The first proposal ( ) is rejected. The second proposal ( ) is drawn uniformly on the ellipse.

Visualized illustration, (c)

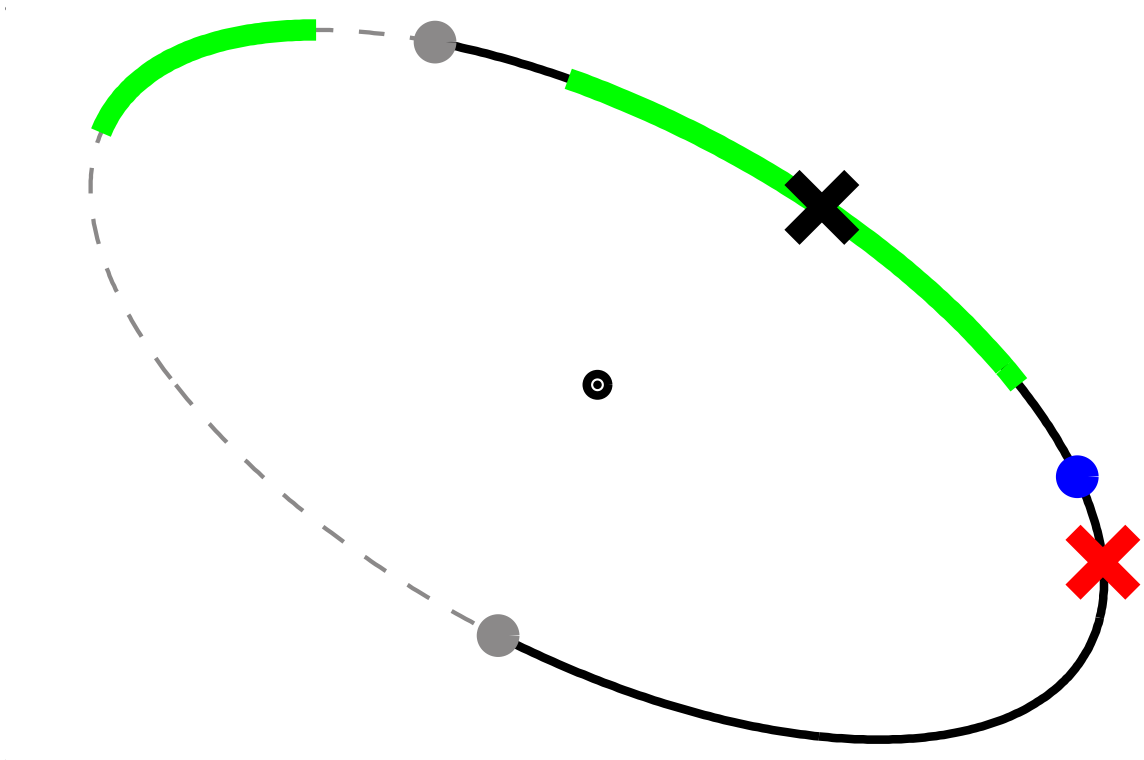

Figure: The first two proposals (●) are rejected. Two proposals divide the ellipse to two brackets. The interval of θ shrinks.
We only draw the third proposal (●) on the part contains ✖.

Visualized illustration, (d)

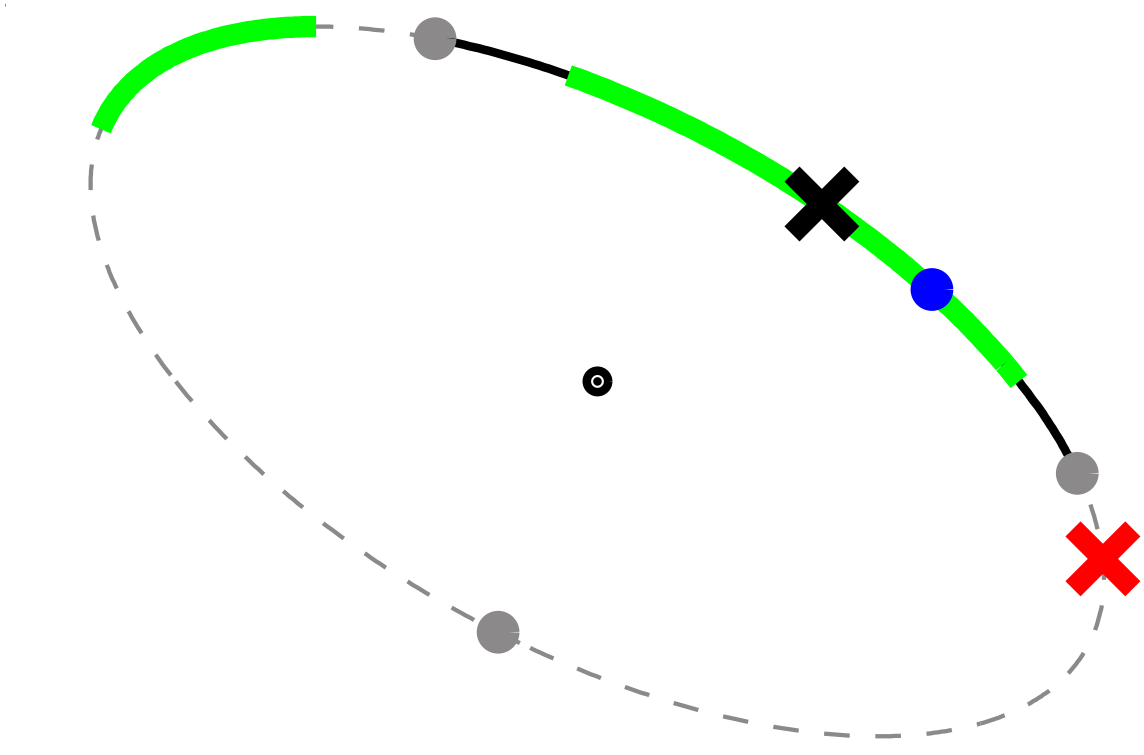

Figure: The third rejected proposal shrinks the interval again. The fourth proposal lies on the green slice, we accept it.

Visualized illustration, (e)

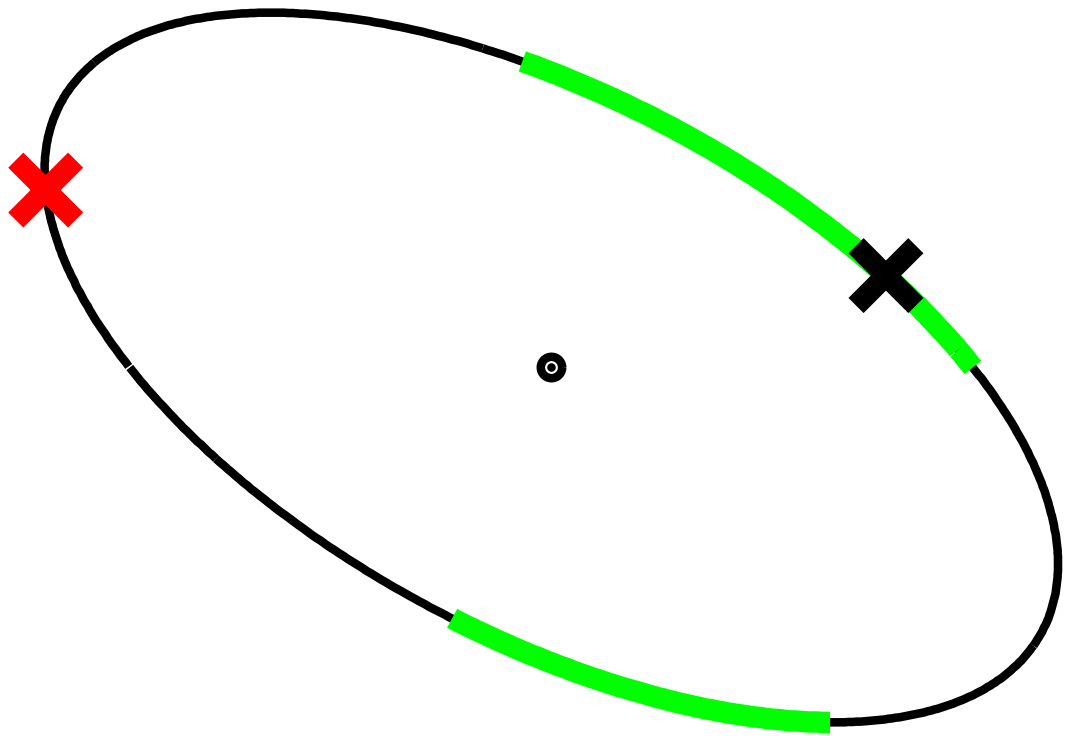

Figure: Second round : The accepted proposal becomes initial value ✖ of the next round. Another auxiliary (✖) value is drawn.
Initial value and auxiliary value define a new ellipse. The acceptance region (green slice) also moves.

(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 tight 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 approximation

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

.

Regression 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 compare with the standard Gibbs sampler : R package `monomvn`, It was written in C++ by Robert B. Gramacy.

We care about

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

Simulations, computing time

n is number of observations, p is number of variables. 15,000 posterior samples are drawn and 3,000 of them are burn-in samples.

Table: MSE and running time comparison

n	p	MSE		Running time (seconds)		
		monomvn	slice	monomvn	slice	ratio
1000	20	0.020	0.020	1.80	0.25	7.20
5000	20	0.004	0.004	7.64	0.23	33.22
1000	50	0.049	0.049	2.36	0.41	5.76
5000	50	0.010	0.010	6.81	0.39	17.46
1000	100	0.105	0.107	6.75	1.49	4.53
5000	100	0.020	0.021	12.28	1.26	9.75
1000	500	0.973	0.985	176.50	8.58	20.57
5000	500	0.110	0.111	213.63	8.51	25.10
5000	1000	0.249	0.249	1214.59	32.76	37.08

Part II : Application to Bayesian IV model

Sparsity versus parsimony

Sparsity

Only a few instruments among a whole batch of candidates are good.
We do not know which ones they are.
Horseshoe prior.

Parsimony

(Hedibert Lopes's yesterday talk, EFaB 311 session.)
Parsimonious correlation structure in $\mathbf{Z}$.
Linear combinations that are strong instruments.
Factor shrinkage prior.

Application to IV model

Here is a way to write IV models

$$\begin{aligned} \mathbf{x} &= \mathbf{Z}\delta + \sigma\epsilon_x \\ \mathbf{y} &= \alpha\sigma\epsilon_x + \mathbf{x}\beta + \xi\epsilon_y. \end{aligned} \tag{3}$$

Note that $\sigma\epsilon_x = \mathbf{x} - \mathbf{Z}\delta$. Writing $\beta^* = (\alpha + \beta)$ and rearranging gives

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

Z being a valid instrument means that δ is the same in both equations and is not the zero vector; this allows us to estimate α.

Application to IV model

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

How Bayesian IV works:

- Regular OLS gives an estimate of β^*, not β (the "structural" parameter).
- The model is a compositional representation where ϵ_x and ϵ_y are *independent.*
- α reflects the impact of unmeasured confounding (y depends on ϵ_x).
- We can estimate α if we can estimate δ, which allows us to solve for β given an estimate of β^*. $\beta = \beta^* - \alpha$.

Gibbs update for β^* and δ^*

To deploy the slice sampler, we write the likelihood as

$$\begin{aligned} \frac{\xi \mathbf{x}}{\sigma} &= \frac{\xi}{\sigma}\mathbf{Z}\delta + \xi\epsilon_x \\ \mathbf{y} &= \mathbf{x}\beta^* - \alpha\mathbf{Z}\delta + \xi\epsilon_y \end{aligned} \tag{6}$$

Therefore, let

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{x} \\ \mathbf{y} \end{pmatrix}, \tilde{\mathbf{Z}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{Z} & 0 \\ -\alpha\mathbf{Z} & \mathbf{x}, \end{pmatrix}, \tilde{\delta} = \begin{pmatrix} \delta \\ \beta^* \end{pmatrix} = \begin{pmatrix} \delta \\ \alpha + \beta \end{pmatrix} \tag{7}$$

and configure the regression as

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{Z}}\tilde{\delta} + \xi\varepsilon \tag{8}$$

Sampling procedures

We construct a big Gibbs sampler for sampling all parameters $(\tilde{\delta}, \alpha, \xi, \sigma)$. Note that β^* is the last entry of $\tilde{\delta}$.

1. Sampling $\tilde{\delta}$ by elliptical slice sampler.

2. Update ξ, σ by inverse-gamma prior.

3. Update α by run regression $\mathbf{y} - \mathbf{x}\beta^* = \alpha(-\mathbf{Z}\delta) + \xi\varepsilon$ with flat prior and horseshoe Metropolis-Hastings adjustment.

Simulation studies

Data is generated from the model

$$\begin{aligned}\mathbf{x} &= \mathbf{Z}\delta + \sigma\epsilon \\ \mathbf{y} &= \underbrace{\alpha(\mathbf{x} - \mathbf{Z}\delta)}_{\text{Variance is } \alpha^2\sigma^2 := \kappa^2} + \underbrace{\mathbf{x}\beta}_{\text{Variance is } \beta^2} + \underbrace{\xi\varepsilon}_{\text{Variance is } \xi^2}\end{aligned} \tag{9}$$

1. Assign 3 parameters : κ^2 and SNR_X^2 between $[0, 1]$ and SNR_Y^2.

2. Let $\sigma = \sqrt{1 - \text{SNR}_X^2}$, $|\alpha| = \sqrt{\kappa^2/\sigma} + N(0, 0.1)$. Assign α a random sign.

3. Draw δ on a radius SNR_X hypersphere.

4. Let $\xi = \sqrt{(\kappa^2 + \beta^2)/\text{SNR}_Y}$.

5. Draw $\beta \sim \text{Unif}[-2\kappa, 2\kappa]$.

6. Draw $\mathbf{Z}, \epsilon, \varepsilon$ from i.i.d. $N(0, 1)$.

Simulation studies

We compute the average length of Anderson-Rubin confidence interval[1] (only finite interval) for 2SLS.

Table: $\mathrm{SNR}_Y^2 = 4$

κ^2	SNR_X^2	0.3			0.5			0.7		
		RMSE	Length	Cover	RMSE	Length	Cover	RMSE	Length	Cover
0.3	2SLS	0.36	10.74	86%	0.26	2.79	92%	0.18	1.34	87%
	Slice	0.15	0.57	97%	0.09	0.41	94%	0.09	0.35	93%
0.5	2SLS	0.46	5.40	76%	0.36	2.62	79%	0.26	1.42	82%
	Slice	0.21	0.75	91%	0.13	0.52	95%	0.12	0.42	94%
0.7	2SLS	0.56	20.02	75%	0.43	5.35	74%	0.29	1.40	73%
	Slice	0.22	0.87	95%	0.14	0.60	98%	0.14	0.51	93%

- Smaller RMSE, length and better coverage than 2SLS.
- Smaller RMSE at strong signal case (large SNR_X^2 and SNR_Y^2)

[1] R package `ivpack`

Simulation studies

Table: $\mathsf{SNR}_Y^2 = 1$

κ^2	SNR_X^2	0.3			0.5			0.7		
		RMSE	Length	Cover	RMSE	Length	Cover	RMSE	Length	Cover
0.3	2SLS	0.38	4.58	98%	0.27	2.82	90%	0.21	1.45	86%
	Slice	0.25	1.01	95%	0.16	0.60	94%	0.12	0.46	93%
0.5	2SLS	0.49	41.21	91%	0.36	2.67	79%	0.26	1.58	78%
	Slice	0.32	1.15	90%	0.18	0.74	98%	0.14	0.61	98%
0.7	2SLS	0.57	7.39	83%	0.42	2.51	71%	0.33	1.59	63%
	Slice	0.50	1.79	92%	0.23	0.91	94%	0.16	0.74	96%

Simulation studies

Table: $\mathrm{SNR}_Y^2 = 0.25$

κ^2	SNR_X^2	0.3			0.5			0.7		
		RMSE	Length	Cover	RMSE	Length	Cover	RMSE	Length	Cover
0.3	2SLS	0.43	5.92	92%	0.31	2.20	85%	0.25	1.53	84%
	Slice	0.45	2.35	89%	0.34	1.04	86%	0.22	0.84	96%
0.5	2SLS	0.56	5.47	87%	0.41	2.41	80%	0.36	1.51	69%
	Slice	0.53	2.54	88%	0.39	1.31	88%	0.30	1.04	92%
0.7	2SLS	0.65	4.95	84%	0.44	2.61	66%	0.42	1.56	63%
	Slice	0.67	3.16	92%	0.34	1.57	97%	0.35	1.18	94%

- No worse than 2SLS when SNR_Y^2 is small.
- Better than 2SLS when signal in first stage is strong. (large SNR_X^2).
- Smaller interval and better coverage than 2SLS.

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.

Y : log(wage).
X : Years of schooling.

We follow the analysis of Hansen and Kozbur (2014)[2]. 509 control variables, consisting of 9 year-of-birth indicators, 50 state-of-birth indicators, as well as the 450 interactions between them.

[2] Instrumental variables estimation with many weak instruments using regularized JIVE. Journal of Econometrics

Returns to schooling data

Three instruments sets are used

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

Returns to schooling results

	2SLS	Post-LASSO	JIVE	RJIVE	FSP	Slice
	3 instruments					
Coeff.	0.1079	0.1150	0.1091	0.1091	0.1098	0.1055
SD	0.0196	0.0205	0.0202	0.0202	0.0207	0.0206
	180 instruments					
Coeff.	0.0928	0.1125	0.1096	0.1062	0.1107	0.1095
SD	0.0097	0.0173	0.0161	0.0157	0.0183	0.0168
	1527 instruments					
Coeff.	0.0712	0.0862	0.0816	0.1067	0.0862	0.0974
SD	0.0049	0.0254	0.5168	0.0171	0.0066	0.0070

OLS : 0.06 (regress Y on X directly.)

Last Slide

Thank you!

Appendix

The "structural" model

The basic linear model is

$$x_i = z_i^t \boldsymbol{\delta} + \epsilon_x$$
$$y_i = \beta x_i + \epsilon_y,$$

where (ϵ_x, ϵ_y) are jointly Gaussian with mean zero and covariance

$$\mathbf{S} := \text{cov} \begin{pmatrix} \epsilon_x \\ \epsilon_y \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}.$$

The reduced form model

The joint distribution of the observables (conditional on z) is

$$
\begin{aligned}
x_i &= \mathrm{z}_i^t \boldsymbol{\delta} + \epsilon_x, \\
y_i &= \mathrm{z}_i^t \boldsymbol{\delta} \beta + \beta \epsilon_x + \epsilon_y.
\end{aligned}
$$

or equivalently

$$
\begin{aligned}
x_i &= \mathrm{z}_i^t \boldsymbol{\delta} + \nu_x, \\
y_i &= \mathrm{z}_i^t \boldsymbol{\delta} \beta + \nu_y,
\end{aligned}
$$

where $\operatorname{cov}\begin{pmatrix} \nu_x \\ \nu_y \end{pmatrix} = \boldsymbol{\Omega} = \mathbf{U}\mathbf{S}\mathbf{U}^t$, with $\mathbf{U} = \begin{pmatrix} 1 & 0 \\ \beta & 1 \end{pmatrix}$.

A reparametrization

Expressing

$$\epsilon_y \mid \epsilon_x \sim \mathsf{N}(\alpha\epsilon_x, \xi^2),$$

gives

$$\alpha = \frac{\sigma_y}{\sigma_x}\rho \quad \text{and} \quad \xi^2 = (1-\rho^2)\sigma_y^2,$$

where $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$.

$$\operatorname{cov}\begin{pmatrix}\nu_x \\ \nu_y\end{pmatrix} = \mathbf{\Omega} = \begin{pmatrix}\sigma_x^2 & (\beta+\alpha)\sigma_x^2 \\ (\beta+\alpha)\sigma_x^2 & (\beta+\alpha)^2\sigma_x^2 + \xi^2\end{pmatrix}.$$

A "two-stage" Bayesian IV model

We now re-express the joint distribution in *compositional form*:

$$
\begin{aligned}
f(x, y \mid \mathrm{z}) &= f(y \mid x, \mathrm{z}) f(x \mid \mathrm{z}) \\
&= \mathsf{N}_{y|x}(x\beta + \alpha(x - \mathrm{z}^t\boldsymbol{\delta}), \xi^2) \times \\
&\quad \mathsf{N}_x(\mathrm{z}^t\boldsymbol{\delta}, \sigma_x^2) \\
&:= \underbrace{f(\mathbf{x} \mid \boldsymbol{\delta}, \sigma_x^2)}_{\text{first stage}} \underbrace{f(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\delta}, \alpha, \beta, \xi^2)}_{\text{second stage}}.
\end{aligned}
$$

This expression brings out an analogy with two-stage least squares (2SLS).

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 $h(\theta^{(t)})$

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

Simulations, Effective Sample Size

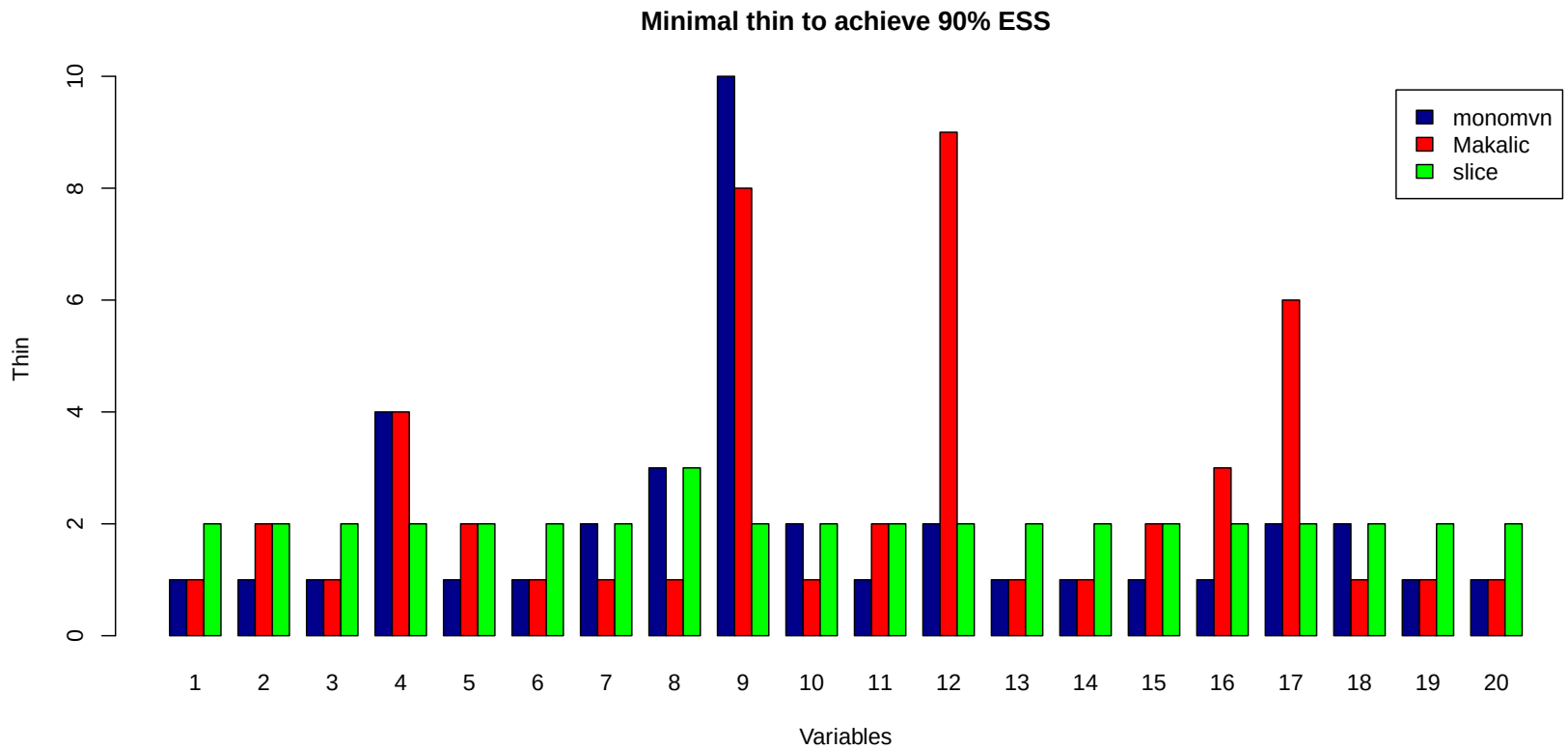

Figure: Minimal among of thinning to achieve 90% ESS for each variable, good data case.

Simulations, Effective Sample Size

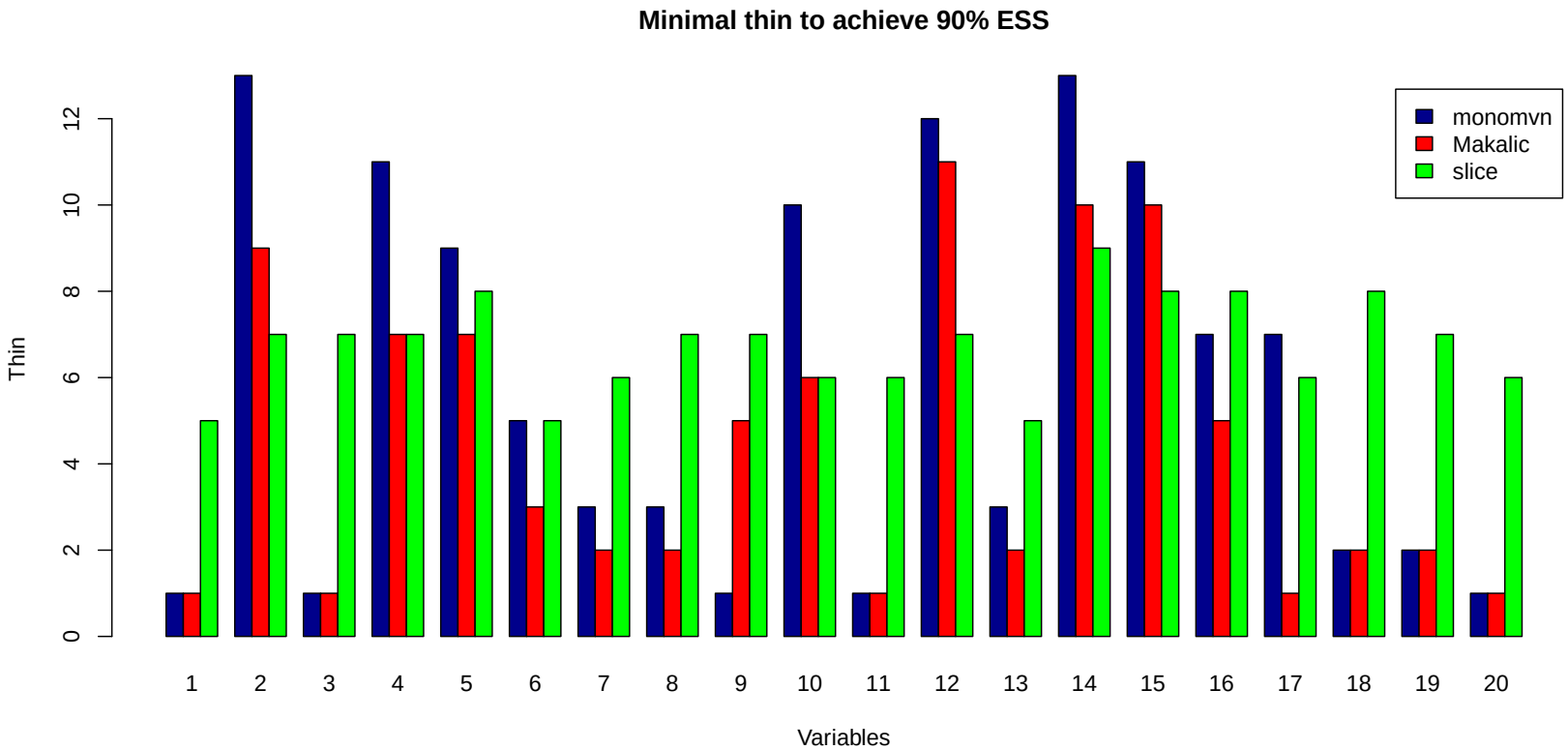

Figure: Minimal among of thinning to achieve 90% ESS for each variable, bad data case

Empirical Example, Effective Sample Size

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

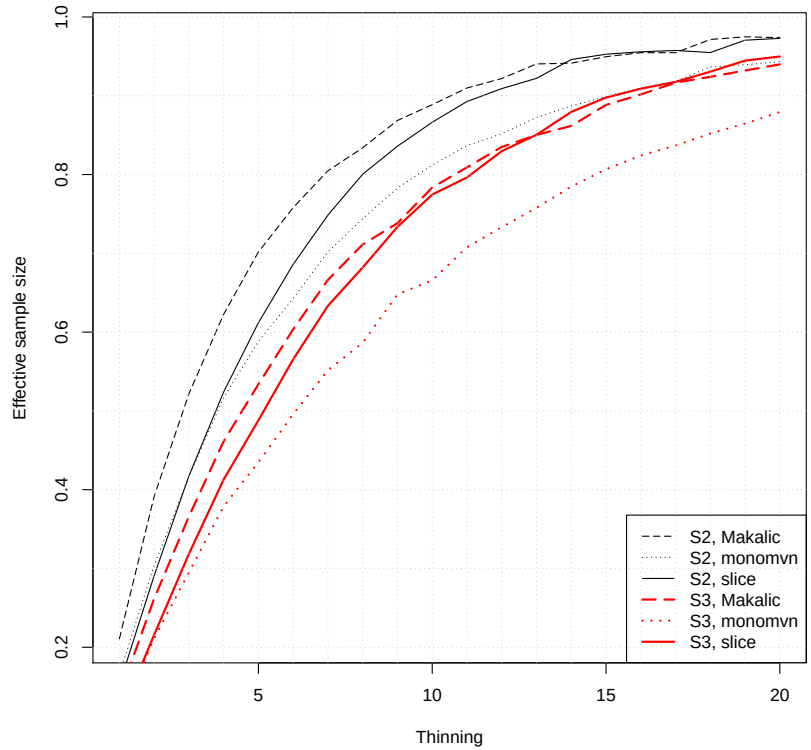

Figure: ESS of Variable `S2` and `S3`, comparing with `monomvn` and Makalic's sampler.

Reference

- 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;

- P. Richard Hahn, Jingyu He & Hedibert Lopes (2016): Bayesian factor model shrinkage for linear IV regression with many instruments, Journal of Business & Economic Statistics, forthcoming.