Elliptical slice sampling for Bayesian shrinkage regression with applications to causal inference

P. Richard Hahn and Jingyu He*
Booth School of Business, University of Chicago
and
Hedibert Lopes
INSPER Institute of Education and Research

August 11, 2016

Abstract

This paper specializes the elliptical slice sampler for Bayesian linear regression models with Gaussian errors and arbitrary shrinkage priors. The only requirement is that the shrinkage prior density function can be evaluated up to a normalizing constant. The new sampler draws all coefficients simultaneously and simulation studies show that it is markedly more efficient than the Gibbs samplers that are typically used for such models. It is demonstrated how to tailor the new sampling scheme to reparametrized linear regressions for causal inference — treatment effect estimation and instrumental variables models.

Keywords: Bayesian linear regression, instrumental variables, treatment effect estimation

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

1 Introduction

Shrinkage estimation refers broadly to the judicious introduction of bias in the service of stabilizing an estimator, resulting in lower average estimation error. In a Bayesian framework this bias is naturally introduced via a prior distribution over the parameters of interest. In a linear regression context, there has been much recent interest in "sparse" models, where the underlying data generating process is believed to arise from a regression vector with very many exactly zero coefficients. Shrinkage priors in this setting refers to priors that place a large fraction of their prior mass on or in a neighborhood around zero. See Hahn and Carvalho [2015] for a thorough review of this literature.

This paper considers the problem of drawing posterior samples from such models. Due to the non-conjugate forms of most shrinkage priors, closed form (e.g., conjugate) posterior distributions are unavailable. At present, most implementations rely on Gibbs samplers to conduct Markov chain Monte Carlo inference. For example, the `R` package `monomvn` [Gramacy, 2010] implements the standard Gibbs sampler for the widely used horseshoe prior Carvalho et al. [2010]. While this approach is relatively simple to implement, it suffers from poor mixing and long run times for many applied data sets where the number of variables p are large (several hundred or thousands). The fundamental difficulty with the Gibbs approach is simply that it must loop through each variable individually.

In this paper, we propose a non-Gibbs sampler based on the elliptical slice sampler of Murray et al. [2010]. These authors focus on sampling from posterior that are proportional to a product of a multivariate normal prior and an arbitrary likelihood. Intuitively, the elliptical slice sampler operates by drawing samples from the normal factor of the posterior and then accepting or rejecting these samples by evaluating the non-normal factor. The starting point of this paper is the observation that the elliptical slice sampler is also suited to the Bayesian normal linear regression case, which has a normal likelihood and an arbitrary shrinkage prior. In this paper, the efficiency of the elliptical slice sampler will be illustrated on linear regression models using the horseshoe prior [Carvalho et al., 2010].

The next section presents the elliptical slice sampler for regression with arbitrary shrinkage priors and reports the results of simulation studies. Section 3 then turns to the application of the new sampler to causal inference problems, for which computational modifications

are necessary to deal with context-specific re-parametrizations. The application of the slice sampler in this context allows standard data sets to be analyzed that were too large to be handled by existing Gibbs-based approaches.

2 Elliptical slice sampler

Unless otherwise noted, random variables (possibly vector valued) are denoted by capital roman letters, matrices are in bold, vectors are in roman font, and scalars are italic. All vectors are column vectors.

Consider the standard Bayesian linear regression model:

$$Y = \mathbf{X}\beta + \epsilon, \tag{1}$$

where Y is an n-by-1 vector of responses, $\mathbf{X}$ is an n-by-p matrix of regressors, β is a p-by-1 vector of coefficients and $\epsilon \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is an n-by-1 vector of error terms. Denote the prior of β as $\pi(\beta)$. The objective is to sample from a posterior expressable as

$$\pi(\beta \mid \mathrm{y}, \mathbf{X}, \sigma^2) \propto f(\mathrm{y} \mid \mathbf{X}, \beta, \sigma^2)\pi(\beta) \tag{2}$$

where $f(\mathrm{y} \mid \mathbf{X}, \beta)$ is a normal likelihood $\mathrm{N}_{\mathrm{y}}(\mathbf{X}\beta, \sigma^2)$. Up to proportionality, $\mathrm{N}_Y(\mathbf{X}\beta, \sigma^2)$ can be regarded as the posterior of β under a flat prior (indicated by the 0 subscript):

$$\begin{aligned} \pi_0(\beta \mid \sigma^2, \mathrm{y}, \mathbf{X}) &\propto \mathrm{N}_Y(\mathbf{X}\beta, \sigma^2) \\ &\propto \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(\mathrm{y} - \mathbf{X}\beta)^T(\mathrm{y} - \mathbf{X}\beta)\right] \\ &\times \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T\mathbf{X}^T\mathbf{X}(\beta - \hat{\beta})\right] \end{aligned} \tag{3}$$

where $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{y}$ is the ordinary least squares (OLS) estimator. Therefore, the slice sampler [Murray et al., 2010] can be applied directly, using $\pi_0(\beta \mid \sigma^2, \mathrm{y}, \mathbf{X})$ as the normal "prior" and $\pi(\beta)$ as the "likelihood".

Elliptical slice sampler for shrinkage regression

For initial value β,

1. Draw $\zeta \sim \mathrm{N}(0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.
2. For $\upsilon \sim \mathrm{Uniform}(0,1)$ define $\ell := \log(\pi(\beta)) + \log(\upsilon)$.
3. Draw angle $\varphi \sim \mathrm{Uniform}(0, 2\pi)$; set $lower \leftarrow \varphi - 2\pi$ and $upper \leftarrow \varphi$.
4. Set $\Delta' \leftarrow \Delta\cos\varphi + \zeta\sin\varphi$ and $\beta' \leftarrow \hat{\beta} + \Delta'$.
5. **while** $\log(\pi(\beta')) < \ell$
   (a) **if** $\varphi < 0$, set $lower \leftarrow \varphi$, **else** set $upper \leftarrow \varphi$.
   (b) Draw angle $\varphi \sim \mathrm{Uniform}(lower, upper)$
   (c) Update $\Delta' \leftarrow \Delta\cos\varphi + \zeta\sin\varphi$ and $\beta' \leftarrow \hat{\beta} + \Delta'$.
6. Set $\Delta \leftarrow \Delta'$ and $\beta \leftarrow \hat{\beta} + \Delta'$.

Murray et al. [2010] verify that the elliptical slice sampler satisfies detailed balance and so generates a valid Markov chain. The algorithm is illustrated in Figure 1. Note that the initial value always lies in the acceptance region, and the region to draw proposals is shrunk to the initial value after every rejection. As a result, acceptance of one proposal is guaranteed. Sampling of σ^2 can be done after sampling β in each iteration, using a conjugate inverse-gamma prior or an arbitrary prior with a Metropolis-Hastings step.

The elliptical slice sampler is flexible because the only requirement is that the prior function $\pi(\beta)$ can be evaluated up to a normalizing constant. In each Monte Carlo iteration, the sampler draws auxiliary sample from multivariate normal distribution once and only needs to draw from a univariate uniform distribution within the while loop. It is not necessary to loop each entry of β, which improves efficiency when dealing with large p cases.

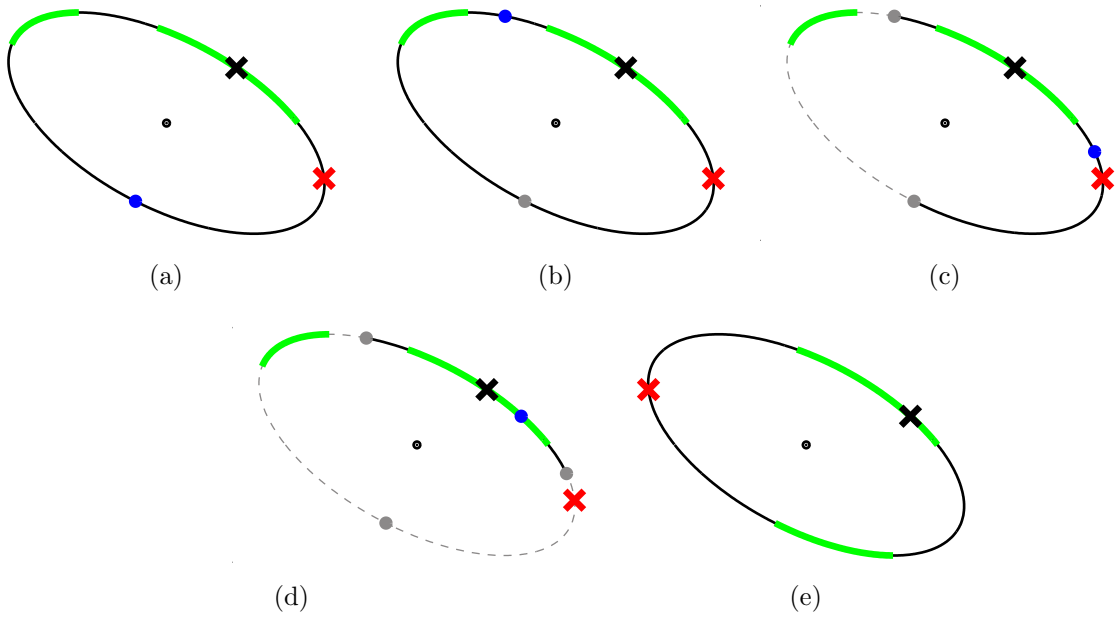

Figure 1: Illustration of the elliptical slice sampler. (a) f (black cross) is the initial value, Step 1 draws auxiliary sample ζ (red cross sign). f and ζ define an ellipse centered at the origin. Step 2: a likelihood threshold defines the slice (green bold line). Step 3 draws an initial proposal (blue dot) which lies out of the slice in this picture. (b) The first proposal defines both edges of $[\theta_{\min}, \theta_{\max}]$, the second proposal is also drawn from the whole ellipse. (c) One edge of the bracket (black line) moves to the last rejected proposal, thus f is still kept inside it. New proposals shrinkage the bracket until one proposal is accepted. (d) this proposal lands on the slice and is returned as $f^{'}$. (e) starts another iteration where $f^{'}$ is the input, and another auxiliary sample (red cross sign) is drawn. Black bracket is shrunk towards initial value after each rejection. Therefore, acceptance of one proposal is guaranteed in each Monte Carlo step. This figure is reproduced from Murray et al. [2010] with modification.

2.1 The horseshoe prior

For the purpose of a concrete demonstration, this paper focuses on the horseshoe prior [Carvalho et al., 2010]. Let $\beta = (\beta_1, ..., \beta_p)$ be a vector of regresion coefficients as above. A horseshoe prior on these coefficients can be expressed as a local scale-mixture of normals

$$\beta \sim N(0, \lambda_0^2 \Lambda^2), \beta_j \sim C^+(0,1), \lambda_0, ..., \lambda_p \sim C^+(0,1), \tag{4}$$

where $\Lambda = \text{diag}(\lambda_1, ...\lambda_p)$ represents the local shrinkage parameters and λ_0 is the global shrinkage parameter. In this exact form, the horseshoe prior lacks a closed-form density function, making it unsuitable for use with the elliptical slice sampler. However, the following bounds provide an excellent approximation which is easy to evaluate,

$$\frac{K}{2} \log\left(1 + \frac{4}{(\beta_j/\lambda_0)^2}\right) < \pi(\beta_j/\lambda_0) < K \log\left(1 + \frac{2}{(\beta_j/\lambda_0)^2}\right) \tag{5}$$

where $K = 1/(2\pi^3)^{1/2}$. Henceforth, when we refer to "the horseshoe prior" we mean the prior with density $\pi(\beta_j/\lambda_0) = \frac{K}{2} \log(1 + 4/(\beta_j/\lambda_0)^2)$. Figure 2 depicts this horseshoe prior. It has a pole at the origin and polynomial tails. The horseshoe prior is a good

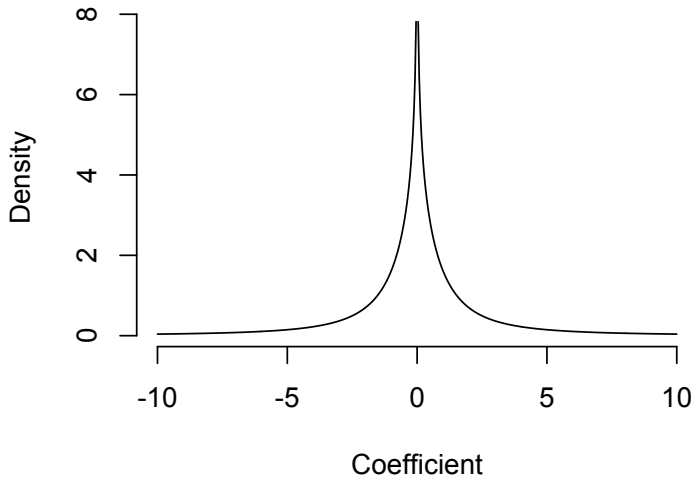

Figure 2: Horseshoe prior, $\pi(\beta_j) = \frac{K}{2} \log(1 + 4/\beta_j^2)$

shrinkage prior for regression coefficients in the sense that it does not require specification

of hyper-parameters and can shrink irrelevant coefficients strongly to zero while keeping the magnitude of relevant coefficients.

By comparison, the standard approach to sampling from the posterior of regressions under horseshoe priors is a Gibbs sampler which samples λ_0 and $(\lambda_1, \ldots, \lambda_p)$ from their full conditionals. This approach is implemented in the `R` package `monomvn` , which needs very long computation time when the number of variables p is large. Makalic and Schmidt [2016] proposes another Gibbs sampler with auxiliary variables (M&S) to improve its performance. However, when the number of variables is large, all the Gibbs samplers are computationally demanding because they need to loop over all coefficients in each interaction. Speed and effective sample size comparisons of the different samplers are given in the following section.

2.2 Simulations

2.2.1 Speed comparison

The `R` package `monomvn` implements the standard Gibbs sampler of horseshoe regression; the appendix of Hahn et al. [2013] provides details of this algorithm. Makalic and Schmidt [2016] propose an alternative Gibb's sampler (M&S). The `monomvn` package and our elliptical slice sampler are written in `C++`, but the original code given in Makalic and Schmidt [2016] is implemented by Matlab. To make our comparison comprehensive, we re-write the M&S sampler in `C++` as well. For each simulation, 13,000 samples are drawn, 3,000 of which are burn-in samples. The true value of β is simulated from the scale-mixture representation horseshoe prior. Table 1 summarizes all the results of simulations. Note the difference of running time between two implementations of the M&S sampler. Our `C++` implementation of M&S sampler is in fact slower than the Matlab code when $p > 500$, likely because Matlab solves linear systems in parallel automatically while our `C++` implementation does not make use of this feature.

The slice sampler enjoys a sizable advantage in terms of computing time compared to the other two methods. For example, when $n = 5000$ and $p = 1000$, our method is about 40 times faster than both competitors. It's noteworthy that the mean squared error (MSE) of the slice sampler is slightly higher than that of the other two samplers when the number of observations n is small, this is likely due to the small difference between our version of

n	p	MSE				Running time (in seconds)			
		M&S		Slice	`monomvn`	M&S		Slice	`monomvn`
		C++	Matlab			C++	Matlab		
50	20	0.611	0.649	0.625	0.614	0.80	2.96	0.24	0.46
100	20	0.230	0.222	0.234	0.230	0.84	3.76	0.23	0.52
500	20	0.038	0.040	0.039	0.038	1.31	4.22	0.24	1.20
1000	20	0.020	0.019	0.020	0.020	1.56	4.85	0.25	1.80
5000	20	0.004	0.004	0.004	0.004	4.53	7.05	0.23	7.64
100	50	0.890	0.848	0.932	0.891	2.94	3.85	0.50	1.57
500	50	0.104	0.098	0.107	0.105	3.28	5.18	0.44	2.09
1000	50	0.049	0.052	0.049	0.049	3.14	5.33	0.41	2.36
5000	50	0.010	0.010	0.010	0.010	7.52	13.76	0.39	6.81
200	100	0.886	0.871	0.912	0.885	11.89	6.43	1.89	6.40
500	100	0.245	0.241	0.251	0.245	10.25	7.88	1.44	5.52
1000	100	0.105	0.108	0.107	0.105	12.20	9.04	1.49	6.75
5000	100	0.020	0.020	0.020	0.020	23.33	28.08	1.26	12.28
1000	500	0.975	0.970	0.986	0.973	262.82	96.51	8.58	176.50
5000	500	0.110	0.112	0.111	0.110	330.79	162.69	8.51	213.63
5000	1000	0.249	0.249	0.249	0.249	1478.59	483.73	32.76	1214.59

Table 1: Mean squared error (MSE) and running time comparison of the elliptical slice sampler, `monomvn` and M&S Gibbs sampler.

the horseshoe prior and the one used to generate the data which the same as is used in the Gibb's sampler implementations. For larger values of n this small discrepancy in the prior becomes increasingly negligible.

2.2.2 Effective sample size

Though the above simulations shows a speed advantage of the elliptical slice sample, one might worry that the raw samples do not tell the whole story and perhaps the new sampler exhibits poor mixing. That is, we care not only about the per-sample speed of the algorithm, but also the quality of sample for the purpose of Monte Carlo inference. To address this concern, here we compare the various samplers on the basis of effective sample size (ESS).

In this simulation study, we again implement horseshoe regression with the elliptical slice sampler and the `monomvn` and M&S Gibb's samplers, with $n = 100$ observations and $p = 20$ predictors. Our data generating process is the same as above. We follow the method in pp. 126 - 127 of Gamerman and Lopes [2006] to compute the ESS. Suppose sample size is N. The effective sample size N_{eff} is

$$N_{\text{eff}} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k}, \tag{6}$$

where $\rho_k = \text{corr}\left(h(\theta^{(0)}), h(\theta^{(t)})\right)$ is the autocovariance of lag k.

The approach is to draw lots of posterior samples by three samplers and compute ESS of each coefficient estimation with different level of thinning. Usually ESS rises with higher thinning, we compare the minimal level of thinning to achieve 90% ESS for each coefficients separately.

Figure 3 and 4 show the minimal thinning of the three samplers, based on two different synthetic data sets. Due to the randomness of simulations, sometimes the synthetic data is easy for all three samplers (in the sense of lower thinning, as shown in figure 3, but two Gibbs samplers get much higher thinning than the elliptical slice sampler on variable 9 and 12. Sometimes the synthetic data is harder to estimate such as shown in figure 4. In this case, the elliptical slice sampler still achieve smaller overall thinning than other two Gibbs samplers. In conclusion, the elliptical slice sampler needs smaller thinning (smaller number of posterior samples) to achieve the same level of effective sample size.

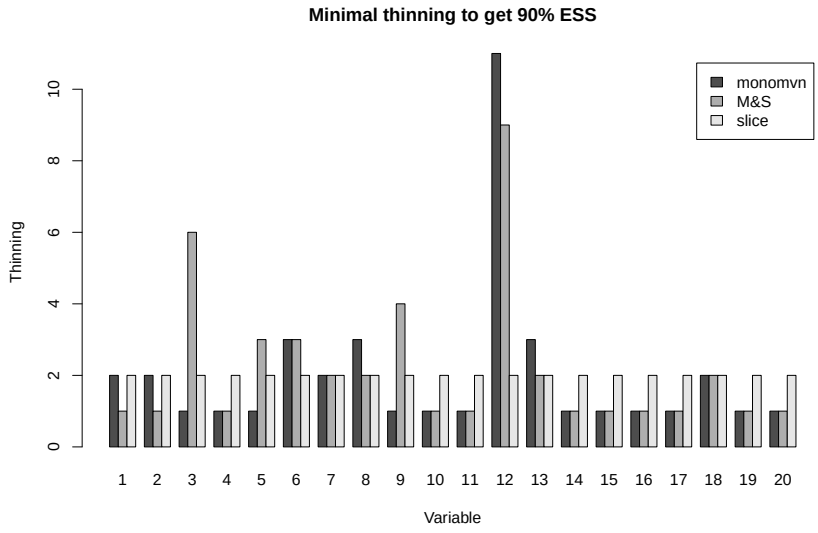

Figure 3: Minimal level of thinning to achieve 90% ESS, easy synthetic data set.

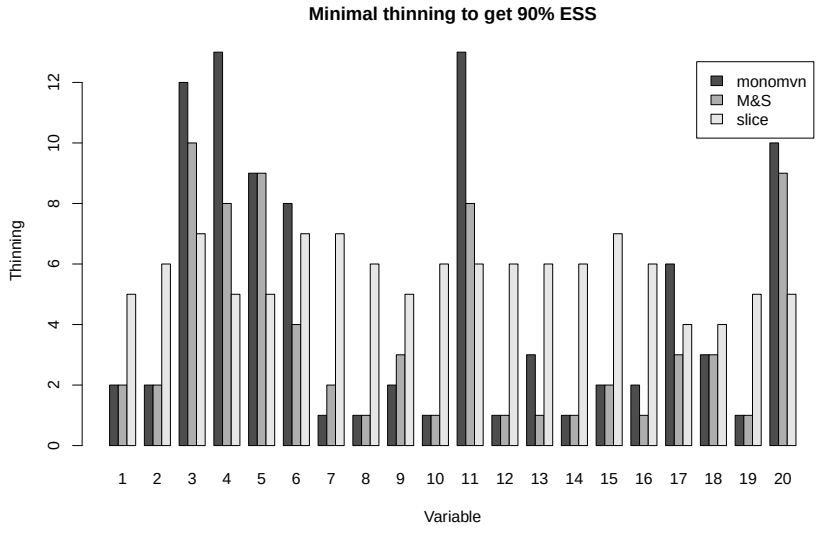

Figure 4: Minimal level of thinning to achieve 90% ESS, hard synthetic data set.

3 Applications to casual inference

The elliptical slice sampler is applicable to many general cases where regularization is desired. We will demonstrate applications to casual inference, in specific, treatment effect estimation and instrumental variable models. In both cases, suppose one has many control variables or instrumental variables, but the minimal set of sufficient control variables or instrumental variables are not known. Therefore regularization plays a key role in estimation of treatment effect or causal effect.

3.1 Treatment effect estimation with no unobserved confounding

This section summarizes the linear model of Hahn et al. [2016a], which is a re-parametrization specifically designed for treatment effect estimation with shrinkage priors.

Consider the linear regression model

$$Y_i = \alpha Z_i + \mathrm{x}_i \beta + \nu_i, \tag{7}$$

where where x_i is a vector of control variables, β is a vector of corresponding coefficients, Z_i is a continuous scalar treatment variable and α is a scalar regression coefficient; assume the dimensions are the same as above. The error terms ν_i are normally distributed with zero mean and unknown variance. In order to estimate α accurately, x_i must include all of the necessary control variables that explain any non-causal association between the treatment Z_i and the response Y_i. In practice, the potential set of control variables is usually large and the minimal sufficient subset is not known. Conventionally, this scenario may be approached by placing a shrinkage prior over β, but as Hahn et al. [2016a] discuss, this "naive" regularization in fact introduces considerable bias (leading to much higher estimation error). To study this phenomenon, the authors study the following two equation linear model (in matrix form)

$$\begin{aligned} &\text{Selection Eq.:} \quad Z = \mathbf{X}\gamma + \epsilon, && \epsilon \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2), \\ &\text{Response Eq.:} \quad Y = \alpha Z + \mathbf{X}\beta + \nu, && \nu \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\nu^2). \end{aligned} \tag{8}$$

In this parameterization, inference concerning α is not impacted by the model for Z_i.

However, under the following simple transformation,

$$\begin{pmatrix} \alpha \\ \beta + \alpha\gamma \\ \gamma \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ \beta_d \\ \beta_c \end{pmatrix}, \tag{9}$$

the model can be written as

$$\begin{aligned} &\text{Selection Eq.:} \quad Z = \mathbf{X}\beta_c + \epsilon, && \epsilon \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2), \\ &\text{Response Eq.:} \quad Y = \alpha(Z - \mathbf{X}\beta_c) + \mathbf{X}\beta_d + \nu, && \nu \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\nu^2). \end{aligned} \tag{10}$$

Now β_c appears in both sampling models and the posterior for α is affected by the model for Z_i. Here, we adapt the elliptical slice sampler to the above model with independent regularization priors over β_c and β_d. In this treatment effect estimation model, we have $\pi(\beta \mid v) = \prod_{j=1}^{p} \log(1+4/(\beta_j/v)^2/v)$. The hyper-parameter v (global shrinkage parameter) is sampled within the Gibbs sampler by a random walk Metropolis update.

Sampling regularized treatment effect linear model

1. Draw $\zeta_1 \sim \mathrm{N}(0, \sigma_\nu^2(\mathbf{X}^T\mathbf{X})^{-1})$ and $\zeta_2 \sim \mathrm{N}(0, \sigma_\epsilon^2(\mathbf{X}^T\mathbf{X})^{-1})$ and defining $\zeta^t = (\zeta_1^t, \zeta_2^t)$.
2. For $\upsilon \sim \text{Uniform}(0,1)$ define $\ell := \log(\pi(\beta_c)) + \log(\pi(\beta_d)) + \log(\upsilon)$.
3. Draw angle $\varphi \sim \text{Uniform}(0, 2\pi)$; set $lower \leftarrow \varphi - 2\pi$ and $upper \leftarrow \varphi$.
4. Set $\Delta' \leftarrow \Delta\cos\varphi + \zeta\sin\varphi$, $\alpha' \leftarrow \hat{\alpha} + \Delta'_1$, $\beta' \leftarrow \hat{\beta} + \Delta'_{2:(p+1)}$, $\gamma' \leftarrow \hat{\gamma} + \Delta'_{(p+2):(2p+1)}$; with $\beta'_c = \gamma'$ and $\beta'_d = \alpha'\gamma' + \beta'$.
5. **while** $\log(\pi(\beta'_c)) + \log(\pi(\beta'_d)) < \ell$
   (a) **if** $\varphi < 0$, set $lower \leftarrow \varphi$, **else** set $upper \leftarrow \varphi$.
   (b) Draw angle $\varphi \sim \text{Uniform}(lower, upper)$
   (c) Update $\Delta' \leftarrow \Delta\cos\varphi + \zeta\sin\varphi$ and $\alpha' \leftarrow \hat{\alpha} + \Delta'_1$, $\beta' \leftarrow \hat{\beta} + \Delta'_{2:(p+1)}$, $\gamma' \leftarrow \hat{\gamma} + \Delta'_{(p+2):(2p+1)}$; with $\beta'_c = \gamma'$ and $\beta'_d = \alpha'\gamma' + \beta'$.
6. Set $\Delta \leftarrow \Delta'$ and $\alpha \leftarrow \hat{\alpha} + \Delta'_1$, $\beta \leftarrow \hat{\beta} + \Delta'_{2:(p+1)}$, $\gamma \leftarrow \hat{\gamma} + \Delta'_{(p+2):(2p+1)}$; with $\beta_c = \gamma$ and $\beta_d = \alpha\gamma + \beta$.

3.1.1 Simulation studies

Here we report a subset of the simulation experiments described in Hahn et al. [2016a]. In these simulations, the new slice sampler is used both in its unmodified form for the "naive" model, as well as the modified form for the two equation regression model for treatment effect estimation.

We control the marginal variance of the treatment and response variables as one, that is, $\text{Var}(Z) = \text{Var}(Y) = 1$.

Define the l_2 norm of the confounding and direct effects as $\rho^2 = ||\beta_c||_2^2$ and $\phi^2 = ||\beta_d||_2^2$. The marginal variances can be decomposed as follows:

$$\begin{aligned} \text{Var}(Z) &= \rho^2 + \sigma_\epsilon^2 \\ \text{Var}(Y) &= \alpha^2(1-\rho^2) + \phi^2 + \sigma_\nu^2 \\ &:= \kappa^2 + \phi^2 + \sigma_\nu^2 \end{aligned} \tag{11}$$

Because marginal variances are fixed to one, $\sigma_\epsilon^2 = 1-\rho^2$ and $\sigma_\nu^2 = 1-\alpha^2(1-\rho^2)-\phi^2$. ρ^2, ϕ^2 and $\kappa^2 := \alpha^2(1-\rho^2)$ measure the percentage of the treatment variance due to confounding, the percentage of the response variance due to the direct impact of the control variables and the percentage of the response variance due to the treatment respectively.

In the following simulation studies, we follow the decomposition of variance in equation (11) and change the strength of the confounding effect ρ^2. $\kappa^2 = \alpha^2(1-\rho^2)$ is fixed, therefore α changes as ρ^2 changes.

Let the number of observations $n = 100, 50$ and the number of control variables $p = 30$ in our simulations. Results of the following variance decomposition parameters are compared: $\{\kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25\}$, $\{\kappa^2 = 0.05, \phi^2 = 0.05, \sigma_\nu^2 = 0.9\}$ and $\rho^2 \in \{0.5, 0.7, 0.9\}$. We show three approaches for estimating the treatment effect as well as an infeasible (oracle) method for comparison: our "two equations" style Bayesian shrinkage approach (new approach), ordinary least squares (OLS), Bayesian shrinkage on the single equation model (7) (naive regularization) and the infeasible "oracle" OLS (OOLS) which runs OLS only using variables with non-zero coefficients.

Table 2 (a) and (b) show results of $\{\kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25\}$ and $n = 100$ and 50 respectively. The direct effect contributes 70% of response variance while the treatment effect derives 5%. Bias, mean square error (MSE), interval length (I.L) and interval coverage

ρ^2		Bias	Cover	I.L.	MSE
0.5	New	-3e-04	0.963	0.335	0.007
	OLS	-0.002	0.951	0.333	0.007
	Naive	-0.077	0.746	0.263	0.012
	OOLS	0.003	0.946	0.292	0.006
0.7	New	0.008	0.964	0.437	0.011
	OLS	0.002	0.944	0.430	0.012
	Naive	-0.156	0.543	0.329	0.035
	OOLS	0.004	0.946	0.377	0.009
0.9	New	-0.004	0.972	0.740	0.029
	OLS	0.005	0.954	0.747	0.035
	Naive	-0.448	0.231	0.478	0.239
	OOLS	0.007	0.946	0.652	0.028

(a) $n = 100, \kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25$.

ρ^2		Bias	Cover	I.L.	MSE
0.5	New	-0.005	0.930	0.518	0.030
	OLS	-0.002	0.944	0.642	0.026
	Naive	-0.087	0.738	0.356	0.022
	OOLS	-0.001	0.952	0.434	0.012
0.7	New	0.006	0.937	0.693	0.034
	OLS	0.005	0.934	0.820	0.048
	Naive	-0.189	0.539	0.403	0.057
	OOLS	-0.002	0.952	0.560	0.020
0.9	New	-0.077	0.959	1.157	0.080
	OLS	-0.016	0.931	1.435	0.140
	Naive	-0.542	0.102	0.487	0.330
	OOLS	-0.003	0.952	0.971	0.059

(b) $n = 50, \kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25$.

ρ^2		Bias	Cover	I.L.	MSE
0.5	New	-0.001	0.933	0.575	0.025
	OLS	-0.004	0.951	0.631	0.026
	Naive	-0.141	0.304	0.175	0.041
	OOLS	0.006	0.946	0.553	0.020
0.7	New	0.004	0.950	0.751	0.037
	OLS	-0.005	0.953	0.816	0.039
	Naive	-0.265	0.134	0.180	0.092
	OOLS	0.008	0.946	0.714	0.033
0.9	New	-0.01	0.939	1.278	0.113
	OLS	-0.002	0.942	1.416	0.135
	Naive	-0.611	0.002	0.184	0.398
	OOLS	0.013	0.946	1.237	0.100

(c) $n = 100, \kappa^2 = 0.05, \phi^2 = 0.05, \sigma_\nu^2 = 0.9$

ρ^2		Bias	Cover	I.L.	MSE
0.5	New	-0.011	0.894	0.819	0.064
	OLS	0.007	0.927	1.204	0.103
	Naive	-0.135	0.349	0.233	0.058
	OOLS	-0.003	0.952	0.824	0.042
0.7	New	-0.028	0.904	1.084	0.105
	OLS	-8e-04	0.938	1.553	0.160
	Naive	-0.275	0.217	0.247	0.116
	OOLS	-0.003	0.952	1.063	0.070
0.9	New	-0.108	0.948	1.813	0.230
	OLS	0.029	0.942	2.671	0.489
	Naive	-0.605	0.015	0.258	0.410
	OOLS	-0.006	0.952	1.842	0.211

(d) $n = 50, \kappa^2 = 0.05, \phi^2 = 0.05, \sigma_\nu^2 = 0.9$.

Table 2: New is new parametrization with horseshoe prior, OLS is ordinary least squares, naive is standard parametrization with horseshoe prior, OOLS is oracle ordinary least squares. MSE is mean square estimation error for the treatment effect and I.L. is interval length for the 95% confidence/credible interval (as appropriate to the method). Results are averages across 100 replications. There are $p = 30$ predictors.

are compared under different levels of confounding effect ($\rho^2 \in \{0.5, 0.7, 0.9\}$). The naive regularization performs poorly in all cases, its bias increases, coverage decreases and MSE explodes when strength of confounding rising. However the interval length is relatively small due to the regularization prior.

Table 2 (c) and (d) show results of the other variance decomposition $\{\kappa^2 = 0.05, \phi^2 = 0.05, \sigma_\nu^2 = 0.9\}$. This is a hard case because the treatment and direct effect only contribute 5% to the response variance separately and the signal-to-noise ratio is low. However our new approach has better estimation relative to OLS for both $n = 100$ and 50. Again, the naive regularization is still under performing.

3.2 Instrumental variables model

Linear instrumental variable (IV) models are widely used to estimate treatment effects of endogenous regressors and thereby perform causal inference. Although usually a single strong instrument is sufficient to identify the treatment effect, researchers are often facing the problem of many weak instruments, whose partial effects are small relative to the residual standard deviation. Transformations, interactions of existing instruments and interactions between instruments and exogenous control variables can generate many candidate instruments easily.

It is well-known that two-stage least squares (2SLS) has large bias when the number of instruments is relatively large to the number of observations [Bekker, 1994, Newey and Smith, 2004]. As a result, shrinkage at the first stage regression is preferred [Carrasco, 2012, Hansen and Kozbur, 2014].

Hahn et al. [2016b] proposes a factor based shrinkage model which implements the horseshoe shrinkage prior by the slice sampler. Here, an alternative parametrization is presented. Inspired by the factorization of the joint likelihood of (x, y) given z in Hahn et al. [2016b],

$$\begin{aligned} f(x, y \mid z) &= f(y \mid x, z) f(x \mid z) \\ &= \mathrm{N}_{y|x}(x\beta + \alpha(x - z\delta), \xi^2) \times \mathrm{N}_x(z\delta, \sigma_x^2) \end{aligned} \tag{12}$$

we write the Bayesian IV model as follows:

$$\begin{aligned} Y &= X\beta + \alpha\sigma\epsilon_X + \xi\epsilon_Y \\ X &= \mathbf{Z}\delta + \sigma\epsilon_X \end{aligned} \tag{13}$$

Y is the response, X is treatment variable and $\mathbf{Z}$ are p dimensional instrumental variables. Note that $\sigma\epsilon_X = X - \mathbf{Z}\delta$, write $\beta^* = \alpha + \beta$ and rearranging gives

$$\begin{aligned} Y &= X\beta^* - \alpha\mathbf{Z}\delta + \xi\epsilon_Y \\ X &= \mathbf{Z}\delta + \sigma\epsilon_X. \end{aligned} \tag{14}$$

β doesn't appear in model (14), regular OLS gives an estimate of β^*. $\mathbf{Z}$ being a valid instrument means that δ is the same in both equations and is not the zero vector, which allows us to estimate α. After estimating β^* and α, it is possible to solve for β by $\beta = \beta^* - \alpha$. α reflects the impact of unmeasured confounding. In specific, Y depends on ϵ_X if α is not zero. The model is a compositional representation where error terms ϵ_X and ϵ_Y are *independent* standard normal.

We work in the parameter space $(\alpha, \xi, \sigma, \beta^*, \delta)$, where $\beta^* = \alpha + \beta$. To conduct inferences concerning β, we transform back, but within the sampler we use the likelihood corresponding to the $(\alpha, \xi, \sigma, \beta^*, \delta)$ parameters.

At the highest level, our sampling approach is a Gibbs sampler, alternating between the five parameters $(\alpha, \xi, \sigma, \beta^*, \delta)$. In fact, β^* and δ are sampled simultaneously. Specifically, note that the model (14) can be written as one multivariate multiple regression given (α, ξ, σ). Let

$$\tilde{\mathrm{y}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathrm{x} \\ \mathrm{y} \end{pmatrix}, \tilde{\mathrm{x}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{Z} & 0 \\ -\alpha\mathbf{Z} & \mathrm{x} \end{pmatrix}, \varepsilon = \begin{pmatrix} \epsilon_{\mathrm{x}} \\ \epsilon_{\mathrm{y}} \end{pmatrix}. \tag{15}$$

The model (14) is reformed as

$$\tilde{\mathrm{y}} = \tilde{\mathrm{x}}\tilde{\delta} + \xi\varepsilon, \tag{16}$$

where

$$\tilde{\delta} = \begin{pmatrix} \delta \\ \beta^* \end{pmatrix} = \begin{pmatrix} \delta \\ \alpha + \beta \end{pmatrix} \tag{17}$$

is a $p+1$ dimensional vector of coefficients. Therefore, it is natural to sample $\tilde{\delta} \mid (\alpha, \sigma, \xi)$ from the reformed model (16) by the elliptical slice sampler. After drawing $\tilde{\delta}^t = (\delta, \beta^*)$, other parameters can be drawn from model (14). In this IV model case, the prior on $\tilde{\delta}$ is

$\pi(\tilde{\delta} \mid \nu) = \prod_{j=1}^{p} \log\left(1 + 4/(\tilde{\delta}_j/v)^2/v\right)$, where the hyper-parameter v can be drawn by a random walk Metropolis update. Following is the entire Gibbs sampling scheme:

Sampling reparametrized IV model

For initial value $(\tilde{\delta}, \alpha, \sigma, \xi)$,

1. Firstly, $\tilde{\delta}$ is drawn by the elliptical slice sampler from $\pi(\tilde{\delta} \mid \alpha, \sigma, \xi, \mathrm{x}, \mathbf{Z}, \mathrm{y})$,

$$\tilde{\mathrm{y}} = \tilde{\mathrm{x}}\tilde{\delta} + \xi\varepsilon,$$

   a regression with prior $\pi(\tilde{\delta}) \propto \pi(\delta)$, the horseshoe prior on the first p entries δ and a flat prior on the last entry β^*.

2. Then we draw (σ, ξ) from model (14) by a conjugate inverse-gamma prior.

   (a) Let δ be the first p entries of $\tilde{\delta}$. (See the definition of $\tilde{\delta}$ in equation (17).) $\mathrm{ssq}_X = ||(X - Z\delta)||_2^2$

   (b) Draw $1/\sigma \sim \mathrm{Gamma}((n + \kappa_X)/2, (\mathrm{ssq}_X + s_X)/2)$.

   (c) Let δ be the first p entries of $\tilde{\delta}$, β^* be the last entry of $\tilde{\delta}$. $\mathrm{ssq}_{\mathrm{y}} = ||\mathrm{y} - \mathrm{x}\beta^* + \alpha\mathbf{Z}^t\delta||_2^2$.

   (d) Draw $1/\xi \sim \mathrm{Gamma}((n + \kappa_{\mathrm{y}})/2, (\mathrm{ssq}_{\mathrm{y}} + s_{\mathrm{y}})/2)$.

3. Finally, update (α, β^*) by regression

$$\mathrm{y} - \mathrm{x}\beta^* = \alpha(-\mathbf{Z}\delta) + \xi\epsilon_{\mathrm{y}}$$

   with flat prior and horseshoe Metropolis-Hastings adjustment on α and a conjugate inverse-gamma prior on ξ.

   (a) $\hat{\alpha} = -(\delta^T\mathbf{Z}^T\mathbf{Z}\delta)^{-1}(\delta^T\mathbf{Z}^T)(\mathrm{y} - \mathrm{x}\beta^*)$. Draw $\alpha = \hat{\alpha} + \xi/\sqrt{\delta^T\mathbf{Z}^T\mathbf{Z}\delta} \times N(0, 1)$.

   (b) To improve mixing, we add an additional step of sampling $\hat{\beta}^*$: $\hat{\beta}^* = (\mathrm{x}^T\mathrm{x})^{-1}\mathrm{x}^T(\mathrm{y} + \alpha\mathbf{Z}^T\delta)$. Draw $\beta^* = \hat{\beta}^* + \sqrt{1/(\mathrm{x}^T\mathrm{x}/\xi^2 + 1)} \times N(0, 1)$.

In the above sampler, $\tilde{\mathrm{x}}$ needs to be updated in each Monte Carlo iteration since it is a function of α, ξ and σ. Due its special structure, all formulas of the above sample can be written in terms of fixed statistics of the observed data: $\mathbf{Z}^T\mathbf{Z}, \mathbf{Z}^T\mathrm{x}, \mathbf{Z}^T\mathrm{y}$ and $\mathrm{x}^T\mathrm{y}$. This strategy proves to be crucial to boost computation speed and save memory; see the appendix for more computational details.

3.2.1 Simulation studies

The simulation study is designed to show the quality of our estimation under different signal strength cases. In specific, we control the relative signal strength from confounding variables and instrumental variables. The data is generated from model (13) as follows.

$$\begin{aligned} X &= \mathbf{Z}\delta + \sigma\epsilon \\ Y &= \underbrace{\alpha(X - \mathbf{Z}\delta)}_{\text{Variance is } \alpha^2\sigma^2 := \kappa^2} + \underbrace{X\beta}_{\text{Variance is } \beta^2} + \underbrace{\xi\varepsilon}_{\text{Variance is } \xi^2} \end{aligned} \tag{18}$$

We scale the marginal variance of X to one and control the marginal variance of Y by manipulating the signal-to-noise (SNR) ratio of the second stage regression. Define the Euclidean $l-2$ norm $\rho^2 = ||\delta||_2^2$, the marginal variances can be decomposed as

$$\begin{aligned} \text{Var}(X) &= \rho^2 + \sigma^2 \\ \text{Var}(Y) &= \kappa^2 + \beta^2 + \xi^2, \end{aligned} \tag{19}$$

where $\kappa^2 = \alpha^2\sigma^2$ represents the signal strength of term $\alpha(X - \mathbf{Z}\delta)$. In the following simulation studies, we set three parameters manually, assign parameters κ^2, ρ^2 between $[0, 1]$ and SNR_Y^2. Let $\sigma = \sqrt{1-\rho^2}$. Fixing the marginal variance of X to be 1, therefore ρ^2 is the percentage of marginal variance due to instrumental variables. δ is drawn on a radius ρ hypersphere. Then $\alpha = \sqrt{\kappa^2/(1-\rho^2)}$, κ^2 represents variance due to confounding $\alpha(Z - X\delta)$ in the second stage regression. The absolute value of β is drawn in the interval $[\kappa/2, 2\kappa]$ uniformly, then β is assigned a positive or negative sign with equal probability 0.5. As a result, in the second stage regression, the signal strength ratio between confounding and treatment effect ranges from 1/2 to 2. Note that due to the decomposition above, the marginal variance of Y is larger than 1 sometimes. To avoid fixing it to 1, we control the signal-to-noise ratio instead, define $\xi = \sqrt{(\kappa^2 + \beta^2)/\text{SNR}_Y}$. Table 3 shows the values of α and ξ under different settings of (κ^2, ρ^2) respectively.

Table 4, 5 and 6 show the simulation results. Root mean squared error (RMSE), credible interval length and coverage are compared. For 2SLS, we follow Anderson and Rubin [1949] to compute the confidence interval, which can be unbounded; we only take average length of finite intervals. In the strong signal-to-noise ratio of Y case (table 4 and 5), our new parametrized model has much smaller root mean squared error (RMSE), credible interval

κ^2 \ ρ^2	0.3	0.5	0.7
0.3	0.6547	0.7746	1.0000
0.5	0.8452	1.0000	1.2910
0.7	1.0000	1.1832	1.5276

(a) $\alpha = \sqrt{\kappa^2/(1-\rho^2)}$

κ^2 \ ρ^2	0.3	0.5	0.7
0.3	0.9972	1.3199	1.8574
0.5	1.2597	1.6763	2.3697
0.7	1.4730	1.9660	2.7861

(b) ξ = Number in the table/SNR_Y

Table 3: Values of (α, ξ) under different settings of κ, ρ

length and better coverage. In weak signal case (table 6, $\rho^2 = 0.3$), we do on par with 2SLS in terms of RMSE and coverage, but our credible interval is still shorter than that of 2SLS.

κ^2 \ ρ^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%

Table 4: Simulation study based on 100 simulated data set, 50 variables. The shrinkage slice sampling method (Slice) is compared with two stages least squares (2SLS) in terms of root mean squared error (RMSE), length of 95% confidence interval (Length) and corresponding cover rates (Cover). In this table $\text{SNR}_Y^2 = 4$

κ^2 \ ρ^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%

Table 5: Simulation study based on 100 simulated data set, 50 variables. The shrinkage slice sampling method (Slice) is compared with two stages least squares (2SLS) in terms of root mean squared error (RMSE), length of 95% confidence interval (Length) and corresponding cover rates (Cover). In this table $\text{SNR}_Y^2 = 1$

κ^2 \ ρ^2		0.3			0.5			0.7		
		RMSE	Length	Cover	RMSE	Length	Cover	RMSE	Length	Cover
0.3	2SLS	0.43	5.92	92%	0.34	2.20	85%	0.25	1.53	84%
	Slice	0.45	2.35	89%	0.33	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	3.15	92%	0.34	1.57	97%	0.35	1.18	94%

Table 6: Simulation study based on 100 simulated data set, 50 variables. The shrinkage slice sampling method (Slice) is compared with two stages least squares (2SLS) in terms of root mean squared error (RMSE), length of 95% confidence interval (Length) and corresponding cover rates (Cover). In this table $\text{SNR}_Y^2 = 0.25$

	2SLS	Post-LASSO	JIVE	RJIVE	FSP	Slice
	A. 3 instruments					
Coeff.	0.1079	0.115	0.1091	0.1091	0.1098	0.1055
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	0.1095
SD	0.0097	0.0173	0.0161	0.0157	0.0183	0.0168
	C. 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

Table 7: Estimation of treatment effects from the new parametrization model and other regularization methods. The table is reproduced from Hahn et al. [2016b] and Hansen and Kozbur [2014].

3.2.2 Empirical example

In this section, we reconsider the well-known analysis of Angrist and Keueger [1991]. The casual impact of schooling on wages is of interest. We analyze data from 1980 U.S. Census on 329,509 men born between 1930 and 1939, construct control variables and instrumental variables as in Hansen and Kozbur [2014], where control for 509 variables including 9 year-of-birth indicators, 50 state-of-birth indicators and 450 interactions between them. Three possible instruments sets are considered. The smallest instruments set contains 3 quarter-of-birth indicators. One can enlarge the 3 instruments set by adding their interactions with the 9 main effects of year-of-birth and 50 main effects of state-of-birth, for a total of 180 instruments. Furthermore, 3 quarter-of-birth indicators and their interactions with all state-of-birth and year-of-birth main effects give a total of 1527 instruments set. For the validation of the quarter-of-birth as instruments, see Angrist and Keueger [1991].

The response variable of interest is the reported log of wage, $y_i = \log(\text{wage}_i)$ and the treatment variable x_i is the reported years of completed education for individual i.

Table 7 compares several different methods, expands the table from Hahn et al. [2016b] and Hansen and Kozbur [2014]. Although all estimates are close in 3 instruments case, it's observed that as more instruments are added, estimates of many estimators go down toward the OLS estimate (0.0673). It is interesting to see how regularization methods mitigate this trend. All five shrinkage approaches get higher estimates than 2SLS using larger instrument sets.

4 Summary

This paper presents an elliptical slice sampler for the purpose of Bayesian linear regression with arbitrary shrinkage priors. The new method is seen to be more computationally efficient than the Gibbs samplers that are routinely used to fit such models. Further we demonstrated the new approach on reparametrized linear models geared specifically towards causal inference; the new sampler allows the causal inference models to be fit to canonical data sets from the applied literature which were too large to be fit by previous computational strategies.

A Computational details for sampling the IV model

A.1 Sampling ζ

The elliptical slice sampler draws $\zeta \sim \mathrm{N}(0_{p\times p}, \xi(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1})$. The matrix inverse $(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}$ must be computed anew at each iteration because it is itself a function of unknown parameters (α, ξ, σ). Fortunately, it is possible to derive $(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}$ in terms of pre-computed quantities, thus avoiding costly matrix inversion.

$$\begin{aligned} \mathrm{y} &= \mathrm{x}\beta^* - \alpha\mathbf{Z}\delta + \xi\varepsilon \\ \frac{\xi\mathrm{x}}{\sigma} &= \mathbf{Z}\delta \cdot \frac{\xi}{\sigma} + \xi\epsilon \end{aligned} \tag{20}$$

Therefore, let

$$\tilde{\mathrm{y}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathrm{x} \\ \mathrm{y} \end{pmatrix}, \tilde{\mathbf{Z}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{Z} & 0_{n\times 1} \\ -\alpha\mathbf{Z} & \mathrm{x} \end{pmatrix}, \tag{21}$$

and configure the regression as

$$\tilde{\mathrm{y}} = \tilde{\mathbf{Z}}\tilde{\delta} + \xi\varepsilon. \tag{22}$$

Where

$$\tilde{\delta} = \begin{pmatrix} \delta \\ \beta^* \end{pmatrix} = \begin{pmatrix} \delta \\ \alpha + \beta \end{pmatrix} \tag{23}$$

$$\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}} = \begin{pmatrix} (\alpha^2 + \frac{\xi^2}{\sigma^2})\mathbf{Z}^T\mathbf{Z} & -\alpha\mathbf{Z}^T\mathrm{x} \\ -\alpha\mathrm{x}^T\mathbf{Z} & \mathrm{x}^T\mathrm{x} \end{pmatrix} := \begin{pmatrix} \gamma A & \alpha b \\ \alpha b^T & c \end{pmatrix} \tag{24}$$

where $\gamma = \alpha^2 + \frac{\xi^2}{\sigma^2}$.

$$\begin{aligned} \Sigma := (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1} &= \begin{pmatrix} \gamma A & \alpha b \\ \alpha b^T & c \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \gamma^{-1}A^{-1} + \alpha^2 k^{-1}\gamma^{-2}A^{-1}bb^TA^{-1} & -\alpha k^{-1}\gamma^{-1}A^{-1}b \\ -\alpha k^{-1}\gamma^{-1}b^TA^{-1} & k^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \gamma^{-1}A^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{k}\begin{pmatrix} \alpha^2\gamma^{-2}A^{-1}bb^TA^{-1} & -\alpha\gamma^{-1}A^{-1}b \\ -\alpha\gamma^{-1}b^TA^{-1} & 1 \end{pmatrix} \end{aligned} \tag{25}$$

where

$$k = c - \alpha^2\gamma^{-1}.$$

Note that

$$A^{-1}b = -(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathrm{x} = -\hat{\beta}_x \tag{26}$$

Plug (26) into (25),

$$\begin{aligned} \Sigma &= \begin{pmatrix} \gamma^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1} & 0_{p\times 1} \\ 0_{1\times p} & 0_{1\times 1} \end{pmatrix} + \frac{1}{k}\begin{pmatrix} \alpha^2\gamma^{-2}\hat{\beta}_x\hat{\beta}_x^T & \alpha\gamma^{-1}\hat{\beta}_x \\ \alpha\gamma^{-1}\hat{\beta}_x^T & 1 \end{pmatrix} \\ &= \begin{pmatrix} \gamma^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1} & 0_{p\times 1} \\ 0_{1\times p} & 0_{1\times 1} \end{pmatrix} + \frac{1}{k}\begin{pmatrix} \alpha\gamma^{-1}\hat{\beta}_x \\ 1 \end{pmatrix}\begin{pmatrix} \alpha\gamma^{-1}\hat{\beta}_x \\ 1 \end{pmatrix}^T. \end{aligned} \tag{27}$$

Suppose

$$LL^T = (\mathbf{Z}^T\mathbf{Z})^{-1} \tag{28}$$

Then

$$\gamma^{-1}\begin{pmatrix} L \\ 0_{1\times 1} \end{pmatrix}\begin{pmatrix} L \\ 0_{1\times 1} \end{pmatrix}^T = \begin{pmatrix} \gamma^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1} & 0_{p\times 1} \\ 0_{1\times p} & 0_{1\times 1} \end{pmatrix} \tag{29}$$

Sampling $\zeta \sim N(0, \xi(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1})$ is simplified as

$$\zeta = \xi \left(\alpha\gamma^{-1}k^{-1/2} \begin{pmatrix} \hat{\beta}_x \\ \gamma\alpha^{-1} \end{pmatrix} \varepsilon_1 + \gamma^{-1/2} \begin{pmatrix} L \\ 0_{1\times 1} \end{pmatrix} \varepsilon_2 \right) \tag{30}$$

where $\varepsilon_1 \sim \mathrm{N}(0,1)$ and ε_2 are i.i.d. $\mathrm{N}(0_{p\times 1}, \mathrm{I}_p)$.

A.2 Computing $\hat{\tilde{\delta}}$

Similarly, the offset (posterior mean) vector $\hat{\tilde{\delta}}$ is a function (α, ξ, σ), but can be expressed in terms of pre-computed quantities. Suppose x and y are column vectors.

$$\tilde{\mathrm{y}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathrm{x} \\ \mathrm{y} \end{pmatrix}, \tilde{\mathbf{Z}} = \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{Z} & 0_{n\times 1} \\ -\alpha\mathbf{Z} & \mathrm{x} \end{pmatrix}, \tilde{\mathbf{Z}}^T = \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{Z}^T & -\alpha\mathbf{Z}^T \\ 0_{1\times n} & \mathrm{x}^T \end{pmatrix} \tag{31}$$

$$\hat{\tilde{\delta}} = (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathrm{y}} \tag{32}$$

Suppose $\hat{\beta}_x = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathrm{x}$, $\hat{\beta}_y = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathrm{y}$, $\hat{\mathrm{x}} = \mathbf{Z}\hat{\beta}_x$, $\hat{\mathrm{y}} = \mathbf{Z}\hat{\beta}_y$

$$(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1} = \begin{pmatrix} \gamma^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1} & 0_{p\times 1} \\ 0_{1\times p} & 0_{1\times 1} \end{pmatrix} + \frac{1}{k} \begin{pmatrix} \alpha^2\gamma^{-2}\hat{\beta}_x\hat{\beta}_x^T & \alpha\gamma^{-1}\hat{\beta}_x \\ \alpha\gamma^{-1}\hat{\beta}_x^T & 1 \end{pmatrix} \tag{33}$$

$$\begin{aligned} \tilde{\mathbf{Z}}^T\tilde{\mathrm{y}} &= \begin{pmatrix} \frac{\xi}{\sigma}\mathbf{Z}^T & -\alpha\mathbf{Z}^T \\ 0_{1\times p} & \mathrm{x}^T \end{pmatrix} \begin{pmatrix} \frac{\xi}{\sigma}\mathrm{x} \\ \mathrm{y} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\xi^2}{\sigma^2}\mathbf{Z}^T\mathrm{x} - \alpha\mathbf{Z}^T\mathrm{y} \\ \mathrm{x}^T\mathrm{y} \end{pmatrix} \end{aligned} \tag{34}$$

Therefore

$$\begin{aligned} \hat{\tilde{\delta}} &= (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathrm{y}} \\ &= \left(\begin{pmatrix} \gamma^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1} & 0_{p\times 1} \\ 0_{1\times p} & 0_{1\times 1} \end{pmatrix} + \frac{1}{k} \begin{pmatrix} \alpha^2\gamma^{-2}\hat{\beta}_x\hat{\beta}_x^T & \alpha\gamma^{-1}\hat{\beta}_x \\ \alpha\gamma^{-1}\hat{\beta}_x^T & 1 \end{pmatrix} \right) \begin{pmatrix} \frac{\xi^2}{\sigma^2}\mathbf{Z}^T\mathrm{x} - \alpha\mathbf{Z}^T\mathrm{y} \\ \mathrm{x}^T\mathrm{y} \end{pmatrix} \\ &= \begin{pmatrix} \gamma^{-1}\frac{\xi^2}{\sigma^2}\hat{\beta}_x - \alpha\gamma^{-1}\hat{\beta}_y \\ 0_{1\times 1} \end{pmatrix} + \begin{pmatrix} k^{-1}\alpha^2\gamma^{-2}\hat{\beta}_x\hat{\beta}_x^T\left(\frac{\xi^2}{\sigma^2}\mathbf{Z}^T\mathrm{x} - \alpha\mathbf{Z}^T\mathrm{y}\right) + k^{-1}\alpha\gamma^{-1}\hat{\beta}_x\mathrm{x}^T\mathrm{y} \\ k^{-1}\alpha\gamma^{-1}\hat{\beta}_x^T\left(\frac{\xi^2}{\sigma^2}\mathbf{Z}^T\mathrm{x} - \alpha\mathbf{Z}^T\mathrm{y}\right) + k^{-1}\mathrm{x}^T\mathrm{y} \end{pmatrix} \\ &= \begin{pmatrix} \gamma^{-1}\frac{\xi^2}{\sigma^2}\hat{\beta}_x - \alpha\gamma^{-1}\hat{\beta}_y + k^{-1}\alpha^2\gamma^{-2}\frac{\xi^2}{\sigma^2}\hat{\beta}_x\hat{\mathrm{x}}^T\mathrm{x} - k^{-1}\alpha^3\gamma^{-2}\hat{\beta}_x\hat{\mathrm{x}}^T\mathrm{y} + k^{-1}\alpha\gamma^{-1}\hat{\beta}_x\mathrm{x}^T\mathrm{y} \\ k^{-1}\alpha\gamma^{-1}\frac{\xi^2}{\sigma^2}\hat{\mathrm{x}}^T\mathrm{x} - k^{-1}\alpha^2\gamma^{-1}\hat{\mathrm{x}}^T\mathrm{y} + k^{-1}\mathrm{x}^T\mathrm{y} \end{pmatrix} \end{aligned} \tag{35}$$

To emphasize, the merit of equation (35) is that $\hat{\tilde{\delta}}$ is expresses as a function parameters (α, ξ, σ) and fixed statisics of the observed data, thus avoiding redundant computations; it is much more efficient than directly computing $(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathrm{y}}$ at each iteration.

References

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, pages 46–63, 1949.

J. D. Angrist and A. B. Keueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

P. A. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

M. Carrasco. A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398, 2012.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, page asq017, 2010.

D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

R. Gramacy. monomvn: Estimation for multivariate normal and student-t data with monotone missingness. *R package version*, pages 1–8, 2010.

P. R. Hahn and C. M. Carvalho. Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

P. R. Hahn, C. M. Carvalho, and S. Mukherjee. Partial factor modeling: predictor-dependent shrinkage for linear regression. *Journal of the American Statistical Association*, 108(503):999–1008, 2013.

P. R. Hahn, C. M. Carvalho, J. He, and D. Puelz. Regularization and confounding in linear regression for treatment effect estimation. Technical report, The University of Chicago Booth School of Business, 2016a.

P. R. Hahn, J. He, and H. Lopes. Bayesian factor model shrinkage for linear iv regression with many instruments. *Journal of Business & Economic Statistics*, 0(ja):1–28, 2016b. doi: 10.1080/07350015.2016.1172968. URL `http://dx.doi.org/10.1080/07350015.2016.1172968`.

C. Hansen and D. Kozbur. Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, 182(2):290–308, 2014.

E. Makalic and D. F. Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, Jan 2016. ISSN 1070-9908. doi: 10.1109/LSP.2015.2503725.

I. Murray, R. P. Adams, and D. J. MacKay. Elliptical slice sampling. In *JMLR Workshop and Conference Proceedings*, volume 9, pages 541–548. JMLR, 2010.

W. K. Newey and R. J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.