

Bayesian Models for Causal Analysis with Many Potentially Weak Instruments

by

Sheng Jiang

Program in Statistical and Economic Modeling
Duke University

Date: _____

Approved:

Surya Tokdar, Supervisor

Merlise Clyde

V. Joseph Hotz

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Program in Statistical and Economic Modeling
in the Graduate School of Duke University
2015

ABSTRACT

Bayesian Models for Causal Analysis with Many Potentially
Weak Instruments

by

Sheng Jiang

Program in Statistical and Economic Modeling
Duke University

Date: _____

Approved:

Surya Tokdar, Supervisor

Merlise Clyde

V. Joseph Hotz

An abstract of a thesis submitted in partial fulfillment of the requirements for
the degree of Master of Science in the Program in Statistical and Economic
Modeling
in the Graduate School of Duke University
2015

Copyright © 2015 by Sheng Jiang
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

This paper investigates Bayesian instrumental variable models with many instruments. The number of instrumental variables grows with the sample size and is allowed to be much larger than the sample size. With some sparsity condition on the coefficients on the instruments, we characterize a general prior specification where the posterior consistency of the parameters is established and calculate the corresponding convergence rate. In particular, we show the posterior consistency for a class of spike and slab priors on the many potentially weak instruments. The spike and slab prior shrinks the number of instrumental variables, which avoids overfitting and provides uncertainty quantifications on the first stage. A simulation study is conducted to illustrate the convergence notion and estimation/selection performance under dependent instruments. Computational issues related to the Gibbs sampler are also discussed.

To my parents.

Contents

Abstract	iv
List of Figures	viii
List of Abbreviations and Symbols	ix
Acknowledgements	xi
1 Introduction	1
2 General Prior and Convergence Results	7
2.1 General Likelihood: the Exponential Family	7
2.2 General Prior	9
2.3 The General Theorem	11
3 The Spike and Slab Prior and Convergence Results	13
3.1 Likelihood: Homoscedastic Gaussian Errors	13
3.2 The Spike and Slab Prior	14
3.3 Posterior Consistency under Spike and Slab Priors	16
3.4 Finite Sample Posterior Inference	25
3.4.1 Likelihood	25
3.4.2 Prior Specification	26
3.4.3 Posterior Distribution: The Gibbs Sampler	27
3.4.4 Computation Issues of The Gibbs Sampler	28

4	Monte Carlo Experiments	30
4.1	Data Generating Process	30
4.2	Prior Specification	31
4.3	Simulation Results	31
4.3.1	Estimation of (β, a_{21}) as $n \rightarrow \infty$	31
4.3.2	Estimation Under Independent/dependent IV	32
5	Discussion	36
A	Appendix	37
A.1	Proof of Theorem 1	37
A.2	Proof of Theorem 3	42
A.2.1	Likelihood Calculation	42
A.2.2	Checking $\pi(\gamma > \bar{r}_n) \leq e^{-4n\epsilon_n^2}$	43
A.2.3	Checking $\pi(\beta > C_n) \leq e^{-4n\epsilon_n^2}$	45
A.2.4	Checking $\pi(\gamma = \gamma_n) \geq e^{-n\epsilon_n^2/8}$	46
A.2.5	Checking $\pi((\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta} \gamma = \gamma_n) \geq e^{-n\epsilon_n^2/8}$	46
A.3	Verify Other Rate Assumptions	48
A.4	Derivation of The Full Conditional of θ_j	49
A.5	Mixing of β and θ	50
	Bibliography	54

List of Figures

4.1	Scatter plots of (β, a_{21}) with independent IV	32
4.2	Scatter plots of (β, a_{21}) with $(n, K_n) = (300, 700)$	33
4.3	θ v.s. $\hat{\theta}$, $(n, K_n) = (300, 700)$	34
4.4	θ v.s. $\tilde{\theta}$, $(n, K_n) = (300, 700)$	35
A.1	Mixing of β , Starting Value=1	51
A.2	Mixing of β at different starting values	52
A.3	Mixing of h_i	53

List of Abbreviations and Symbols

Symbols

\mathbb{R}	Real number.
\ln	Natural logarithm.
z	Instrumental variable vector.
x	Endogenous explanatory variable.
y	Response variable.
ς	Nuisance parameter.
n	Sample size.
K_n	Length of z , the number of instrumental variables.
\bar{r}_n	A cap on the number of instrumental variables selected.
\mathcal{P}_n	A sequence of sets of probability densities.
ψ	A matrix of parameters defined in Chapter 2.1.
Ψ	Some nondecreasing convex function.
$\ \cdot\ _\Psi$	Orlicz norm with respect to function Ψ .
$ \cdot $	Absolute value or ℓ_1 norm.
$d(\cdot, \cdot)$	Hellinger distance between two probability densities.
$a_n > b_n$	a_n is of higher order than b_n : $a_n/b_n \rightarrow \infty$.
$a_n \gtrsim b_n$	a_n is of no lower order than b_n : $a_n/b_n \rightarrow \infty$ or $a_n/b_n \rightarrow \text{constant}$.
$X \perp\!\!\!\perp Y$	Random variable X is independent of random variable Y .

Abbreviations

IV	Instrumental Variable
TOLS	Two Stage Least Square
LIML	Limited Information Maximum Likelihood
Lasso	Least Absolute Shrinkage and Selection Operator
MCMC	Markov Chain Monte Carlo
EM	Expectation Maximization (algorithm)
MSE	Mean Square Error
KL	KullbackLeibler

Acknowledgements

I would like to express my special appreciation and thanks to my supervisor professor Surya Tokdar who has been a tremendously invaluable mentor for me. I would not have finished this thesis without his support. Through working on the thesis, I obtain a deeper understanding of Bayesian statistics and statistical modeling in general. I would also like to thank my committee members, professor Merlise Clyde, professor Joe Hotz, for serving as my committee members and letting my defense be an enjoyable moment. I also want to thank professor David Dunson who introduced Bayesian statistics to me, professor Merlise Clyde who taught me (Bayesian) linear models, professor Mike West who taught me probability models and MCMC, professor Jianguo Liu who taught me math in function spaces, professor Wenxin Jiang who shared his recent work with me.

Special thanks to my family. Words simply cannot show how grateful I am to my parents for their solid support. I would also like to thank all of my friends who have made my time at Durham wonderful and unforgettable.

1

Introduction

In employing the linear model $Y = X\beta + \varepsilon$ to make causal inference, the unobserved error ε is often correlated with explanatory variables X , i.e. $EX\varepsilon \neq 0$, in which case the ordinary least square (OLS) estimator is inconsistent and biased. Causal interpretation of the ordinary least square estimates of β is not plausible. The explanatory variables correlated with the unobserved error are called endogenous variables and the ones not correlated with the error are called exogenous variables.

The common rescue is to find a instrumental variable(s) Z , variable(s) that is (are) correlated with the endogenous variable but not correlated with the error term:

$$\begin{aligned} \text{cov}(X, Z) &\neq \mathbf{0} && \text{(first stage relevance)} \\ EZ\varepsilon &= \mathbf{0} && \text{(exclusion restriction)}. \end{aligned}$$

The key idea behind employing instrumental variables (IV) is quasi-experimental, treating the IV as a random treatment assignment scheme that only changes the endogenous explanatory variable and affects the response variable *only* through the instrumented explanatory variable (without affecting the response through other explanatory variables). IV regression makes it possible to draw causal inference with only observational data. (See Angrist et al. (1996))

With IV at hand, the model becomes

$$\begin{aligned} x &= z^T \theta + \varepsilon_1 && \text{(the first stage)} \\ y &= x^T \beta + \varepsilon_2 && \text{(the structural equation),} \end{aligned}$$

where $z \perp (\varepsilon_1, \varepsilon_2)$ and $\text{corr}(\varepsilon_1, \varepsilon_2) \neq 0$. Under this model, consistent estimators can be constructed. For instance, Two-Stage Least Square (TSLS) estimator and Limited Information Maximum Likelihood (LIML) estimator. These estimators are consistent and asymptotically Normal when the strength of IVs is not weak and the sample size is relative large to the number of instruments.¹ However, when we may be able to find a class of instruments by constructing interaction terms or a series of dummy variables, the constructed IV may be weak and we *a priori* do not know which instruments are weak.

When the strength of the available instruments are weak (weak instruments problem) or the number of IV is large relative to sample size (many instruments problem), the asymptotic properties are different from those of a handful number of strong IV. The theoretical aspects of TSLS/LIML in many or weak instruments problems are well understood and robust inference procedures are proposed in Stock et al. (2002); Stock and Yogo (2005); Bekker (1994); Chao and Swanson (2005); Hansen et al. (2008); Anderson et al. (2010); Hausman et al. (2012). The number of instruments along this line of literature is assumed to be growing at the same rate of the sample size, that is, the number of IV is large but still smaller than the sample size.

When the number of instruments is equal to or greater than sample size, neither TSLS estimator nor LIML estimator exists due to rank deficiency of $Z^T Z$ whose inverse is a key ingredient of the projection matrix onto the column space of IV.

¹ There are multiple measures of the strength of IV, e.g., concentration parameter $\mu^2 = n\theta^T \Sigma_z \theta / \sigma_{22}^2$, Wald statistic $W = \hat{\theta}^T Z^T Z \hat{\theta} / \hat{\sigma}_{22}^2$ and first stage F -statistic $W / \dim(Z)$. The variance-covariance matrix of IV is denoted as Σ_z and $\hat{\theta}$ denotes OLS estimates of θ . Though different at appearance, these statistics essentially measure the correlation between the IV and the endogenous variable(s).

A number of studies propose procedures to draw robust inference from many instruments where the number of instruments is allowed to be larger than the sample size. (Belloni et al., 2012; Caner, 2009; Bai and Ng, 2010; Kapetanios and Marcellino, 2010) The idea of their procedures is in two steps: first do variable selection or dimension reduction in the first stage and then use the estimated instruments/information to do inference in the structural equation. Lasso based procedures suggested in Belloni et al. (2012) have attractive theoretical properties: the Lasso based IV estimator is \sqrt{n} -consistent and asymptotically Normal; the estimator is semi-parametrically efficient under homoscedasticity; the procedure allows for imperfect model selection. Despite these good features, to obtain appropriate uncertainty quantifications on the first stage coefficients within the lasso framework, one needs the theory proposed in Lockhart et al. (2014) to do hypothesis testing on the significance of the instruments selected by lasso. In empirical studies, researchers may be interested in the parameters in the first stage. For instance, due to the quasi-experimental nature of IV methods, they may be interested in if the estimated first stage coefficients are consistent with their intuition and identifying more important IV(s).

In contrast, Bayesian approaches offer natural uncertainty quantifications on the variables selected in a unified framework. (Kyung et al., 2010) However, many instruments problem under Bayesian framework are less explored ². The most relevant paper is Chamberlain and Imbens (1996) which proposes a shrinkage estimator, using *all* the instruments at hand.

The spike and slab prior performs both shrinkage and selection simultaneously (George and McCulloch, 1993, 1997) and is not entirely new to economists (Varian, 2014). Chamberlain and Imbens (1996) can be considered as a special case of the spike and slab prior without the spike part. This type of prior directly addresses

² Some popular textbooks in Bayesian econometrics discuss the identification problem induced by the weak instrument(s) but do not offer concrete solutions. (e.g. see Lancaster (2004); Rossi et al. (2005))

the many instruments problem and has clear advantages over prior specifications like Drèze prior (Drèze, 1977; Dreze and Richard, 1983), Jeffreys’s prior (Chao and Phillips, 1998), natural conjugate prior (Hoogerheide et al., 2007), etc.

This paper investigates asymptotic properties of Bayesian IV models that encourage shrinkage/selection from a large number of potentially weak instruments. The “a large number of potentially weak instruments” is coined to express first stage sparsity. The sparsity of the first stage requires that there are infinitely many weak instruments and is a nice approximation to reality: strong IV(s) is(are) rare. The existence of some relatively strong instrument(s) is necessary for the identification of the coefficients on the endogenous variable.

The number of IV grows with the sample size and is allowed to be much larger than the sample size. With some sparsity condition on the coefficients of the first stage, we first characterize a general prior specification where the posterior consistency of the parameters is established and calculate the corresponding convergence rate. In particular, we show posterior consistency holds for a class of spike and slab priors on the many potentially weak instruments. Simulation results show that moderately informative priors are able to select a suitable amount of instruments and the estimation accuracy of the coefficients on the endogenous variables becomes better as sample size goes to infinity allowing for the number of IV going to infinity.

This paper’s contribution is twofold: 1) establishes a theorem to investigate posterior consistency and convergence rate under general prior specifications with likelihood in the exponential family; 2) shows posterior consistency of a class of spike and slab priors with homoscedastic Gaussian error and calculates corresponding convergence rate.

A Brief Review of Bayesian IV Models

In the Bayesian IV literature, one eminent strand of thinking focuses on choos-

ing different types of priors. Drèze (1977) proposes a diffuse prior that is invariant to structural form and reduced form of the model and Dreze and Richard (1983) summarizes the Bayesian simultaneous equations model. Chao and Phillips (1998) proposes the Jeffreys prior in the limited information analysis of the simultaneous equations model. Kleibergen and Zivot (2003) explore the priors that produce posterior distributions that are the same as the sampling distribution of classical TSLS and LIML estimators. Hoogerheide et al. (2007) propose natural conjugate priors that are more informative than Jeffreys prior and Drèze (1977)'s prior. Lopes and Polson (2014) propose a Cholesky based prior on the error terms instead of a inverse Wishart.

However, these priors do not address the problem of selecting from many potentially weak instruments. One exception is that Chamberlain and Imbens (1996) impose a hierarchical structure on the coefficients of the instruments and find this modeling strategy efficiently combines many possibly weak instruments.

Apart from choosing priors, Koop et al. (2012) takes the model averaging approach and designs a reversible jump MCMC algorithm to explore the model space. Instead of working with homoscedastic Gaussian errors, Conley et al. (2008) takes the Bayesian semi-parametric approach, allowing for flexible distributions in the errors. They show that when errors are non-normal, their procedure is more efficient than standard Bayesian or classical methods. Instead of studying the weak instrument, Conley et al. (2012); Chan and Tobias (2014) focus on the exclusion restriction and study the consequences of plausibly exogenous case.

Paper Structure

In the thesis, Chapter 2 presents a general theorem that is useful in establishing posterior consistency of Bayesian IV models and calculating convergence rates. Chapter 3 first presents posterior consistency and convergence rates results for a

class of spike and slab priors and then discusses which prior specifications have theoretical guarantee from the theorem and computation issues with respect to the spike and slab prior. A Gibbs sampler is derived to sample from posterior distribution. Chapter 4 illustrates the convergence and related computation issues mentioned in Chapter 3 through a simulation example. Chapter 5 concludes the paper with a brief discussion on interesting directions that can be done in the future.

2

General Prior and Convergence Results

2.1 General Likelihood: the Exponential Family

The model with homoscedastic Gaussian error is

$$\begin{aligned}x &= z^T \theta + \varepsilon_1 \quad (\text{the first stage}) \\y &= \beta x + \varepsilon_2 \quad (\text{the structural equation}),\end{aligned}$$

where the K_n dimensional z is independent of $(\varepsilon_1, \varepsilon_2)$, $[\varepsilon_1, \varepsilon_2]^T \sim N(\mathbf{0}, \Sigma)$ and

$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix}$. The endogeneity problem arises when $\sigma_{12} \neq 0$ and in this case,

OLS estimator of the linear regression $y = x\beta + \varepsilon_2$ is inconsistent and biased.

Denote $z^T \theta$ as h , then $\begin{bmatrix} x - h \\ y - h\beta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$ and its distribution is a bivariate Normal distribution

$$\begin{bmatrix} x - h \\ y - h\beta \end{bmatrix} \sim N(0, \tilde{\Sigma}),$$

where $\tilde{\Sigma} = \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \Sigma \begin{bmatrix} 1 & \beta \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ a_{21} + \beta & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{2|1}^2 \end{bmatrix} \begin{bmatrix} 1 & a_{21} + \beta \\ 0 & 1 \end{bmatrix}$. Suppose

the instruments are entirely useless, i.e. $\theta = 0$, then $h = 0$. With observations

$\{(x_i, y_i, z_i)\}_{i=1}^n$, we can only estimate components of $\tilde{\Sigma}$ and cannot identify β or a_{21} from the sum $a_{21} + \beta$. To identify β requires at least one θ_j , $j = 1 : K_n$, is not zero, such that $h \neq 0$ for all z 's and, hence, the mean of $[x, y]^T$ can be used to identify β and a_{21} .

So the likelihood is

$$\begin{aligned} f(x, y|h, \varsigma) &= (2\pi)^{-1} |\tilde{\Sigma}|^{-1/2} e^{-\frac{1}{2} \begin{bmatrix} x-h \\ y-h, \beta \end{bmatrix}^T \tilde{\Sigma}^{-1} \begin{bmatrix} x-h \\ y-h, \beta \end{bmatrix}} \\ &= (2\pi)^{-1} (\sigma_{11}^{-2} \sigma_{21}^{-2})^{1/2} e^{-\frac{1}{2} \{ \sigma_{11}^{-2} (x-h)^2 + \sigma_{21}^{-2} (y - (a_{21} + \beta)x + a_{21}h)^2 \}} \end{aligned}$$

The Gaussian error model is essentially an exponential family model. It motivates the regular multivariate exponential family:

$$f(x, y|h, \varsigma) = e^{a(h, \varsigma) \cdot T(x, y) + b(h, \varsigma) + c(x, y)}$$

where $h = z^T \theta$, ς denotes the vector of m nuisance parameters, set $J = m + 1$, $a : \mathbb{R}^J \rightarrow \mathbb{R}^J$, $T(x, y)$ is J dimensional vector of linearly independent functions of x and y , and $b : \mathbb{R}^J \rightarrow \mathbb{R}$.

We are selecting useful variables from the z vector. Variable selection is imposed on z so in showing the consistency results, all other parameters can be treated as nuisance parameters even though β is of primary importance in IV models.

Assume $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are C^1 , i.e., continuously differentiable. This is assumption is mild because it is typically satisfied for many distributions in the exponential family, e.g., Gaussian, Poisson, etc..

Let $a_i(h, \varsigma)$, $b_i(h, \varsigma)$ denote the partial derivative with respect to i^{th} component of (h, ς) , $a_i^j(h, \varsigma)$ denotes the j^{th} component of the partial derivative where $i = 1 : J$ and $j = 1 : J$.

Let $\psi(h, \varsigma)$ denote the expectation of $T(x, y)$ with respect to parameter (h, ς) :

$$\psi(h, \varsigma) = E_{(h, \varsigma)} T(x, y) = \int p_{h, \varsigma} T(x, y) \nu_x(dx) \nu_y(dy),$$

where $p_{h,\varsigma}$ denotes the probability density with respect to parameter (h, ς) . $\psi^j(h, \varsigma)$ denotes the j^{th} component of the $J \times 1$ vector $\psi(h, \varsigma)$.

As $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ is assumed to be C^1 , for $i = 1 : J$,

$$a_i(h, \varsigma) \cdot \psi(h, \varsigma) + b_i(h, \varsigma) = 0.$$

Denote $A(h, \varsigma) \equiv (a_1, a_2, \dots, a_{m+1})^T$ and $B(h, \varsigma) \equiv (b_1, b_2, \dots, b_{m+1})$, then the equations can be written as

$$A(h, \varsigma) \psi(h, \varsigma) + B(h, \varsigma) = 0.$$

Solving the linear system yields

$$\psi(h, \varsigma) = -A(h, \varsigma)^{-1} B(h, \varsigma),$$

provided the invertibility of $A(h, \varsigma)$.

For some known distributions, $\psi(h, \varsigma)$ is known and solving the linear system may be saved. But it is necessary to check the invertibility of $A(h, \varsigma)$ if we need to solve the linear system to get $\psi(h, \varsigma)$.

2.2 General Prior

Condition O and Condition N characterize the general prior such that prior probability over some neighborhood of true density is not too small and the prior probability outside an expanding “sieve” is not large. These two conditions are sufficient to establish posterior consistency. Throughout the paper, $\varepsilon_n \in (0, 1]$ is a sequence of positive real numbers such that $\varepsilon_n^2 n \rightarrow \infty$.

Condition O has two parts: the first part is on the growth rates of the number of IV, K_n , the cap on the number of useful instruments, \bar{r}_n , a measure of the complexity of the likelihood function, $D(r, R)$, and ε_n ; the second part is on the prior probabilities of the region outside an expanding set.

Condition O The outside condition requires there exist some $C_n > 0$ and $\bar{r}_n \in [1, K_n)$ such that

(a) $\bar{r}_n \ln(1/\varepsilon_n^2) < n\varepsilon_n^2$, $\bar{r}_n \ln K_n < n\varepsilon_n^2$, $\bar{r}_n \ln D(\bar{r}_n, C_n) < n\varepsilon_n^2$, where

$$D(r, R) = 1 + (r + m)R \left[\max_{i,j} \sup_{(h,\varsigma) \in B_{r,R}^\infty} |\alpha_i^j(h, \varsigma)| \right] \left[\max_j \sup_{(h,\varsigma) \in B_{r,R}^\infty} |\psi^j(h, \varsigma)| \right],$$

and $B_{r,R}^\infty$ denotes the ℓ_∞ ball with the radius characterized by (r, R) such that $(h, \varsigma) \in B_{r,R}^\infty$ iff $|h| \leq rR$ and $\|\varsigma\|_\infty \leq R$, i.e., $|\varsigma_j| \leq R$ for all $j = 1 : m$.

(b) for large n ,

- $\pi(|\gamma| > \bar{r}_n) \leq e^{-4n\varepsilon_n^2}$.
- $\forall \gamma$ with $|\gamma| \leq \bar{r}_n$, $\forall j \in \gamma$, $\pi(|\theta_j| > C_n|\gamma|) \leq e^{-4n\varepsilon_n^2}$.

Alternatively, when the prior specification is hierarchical, $\forall \gamma$ with $|\gamma| \leq \bar{r}_n$,

$$1 - \pi\left(\bigcap_{j=1:|\gamma|} \{|\theta_j| < C_n\} \mid \gamma\right) \leq \bar{r}_n e^{-4n\varepsilon_n^2}.$$

- $\pi(|\varsigma_j| > C_n) \leq e^{-4n\varepsilon_n^2}$ for all $j = 1 : m$.

Condition N also has two parts: we can construct a sequence of models such that the models in the sequence capture important instruments and the prior probabilities of some neighborhood of the truth is not too small.

Condition N The neighborhood condition requires a sequence of models γ_n exists such that

(a) $\sum_{j \notin \gamma_n} |\theta_j^*| < \varepsilon_n^2$, where θ_j^* denotes the truth of θ_j .

(b) for any sufficiently small η , there exists N_η such that for all $n > N_\eta$,

- $\pi(\gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$;

- $\pi((\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta} | \gamma = \gamma_n) \geq e^{-n\epsilon_n^2/8}$, where $M_{\gamma_n, \eta} = (\theta_j^* \pm \eta\epsilon_n^2/|\gamma_n|)_{j \in \gamma_n} \cup (\varsigma_j^* \pm \eta\epsilon_n^2)_{j=1}^m$, and ς_j^* denotes the truth of ς_j .

2.3 The General Theorem

The posterior consistency is in the sense of Hellinger distance which is a metric on probability densities. Denote Hellinger distance between the density with respect to $(\theta_\gamma, \gamma, \varsigma)$ and the true density $p^*(y, x, z) = f(y, x|z, \theta_{\gamma^*}^*, \gamma^*, \varsigma^*) f_z(z)$ as

$$d(p, p^*)^2 = \int \int \int [p(y, x|z, \theta_\gamma, \gamma, \varsigma) f_z(z) - p^*(y, x, z)]^2 \nu_y(dy) \nu_x(dx) \nu_z(dz)$$

Theorem 1 (Convergence rate under general prior). *The covariates of the model are bounded $|z_j| \leq 1$ for all $j = 1 : K_n$. Suppose the true coefficients on z are absolutely summable: $\lim_{n \rightarrow \infty} \sum_{j=1}^{K_n} |\theta_j^*| < \infty$, where K_n is nondecreasing in n .*

If the prior specification satisfies condition O and condition N, then for n large enough,

$$E^*[\pi(d(p, p^*) > 4\epsilon_n | D^n)] \leq 4e^{-n\epsilon_n^2/2}.$$

The expectation $E^*[\cdot]$ is taken under the true probability density with respect to true parameters $(\theta_{\gamma^*}^*, \gamma^*, \varsigma^*)$, i.e., calculating the average of the posterior probabilities by repeatedly sampling D^n from p^* .

We prove Theorem 1 using Proposition 1 in Jiang (2007). Theorem 1 generalizes the one-dimensional conditions in Jiang (2007) to multivariate conditions hence shows a multivariate version of the general theorem in the paper. The proof of Theorem 1 follows the same idea as in Jiang (2007) and adapts to the multivariate conditions. The details of the proof are in Appendix A.1.

The theorem is general and can be applied to prove posterior consistency under specific classes of prior specifications. But also notice that this theorem provides only

sufficiency for the theoretical guarantee on the posterior consistency and convergence rate. We may have examples that do not satisfy the conditions but are consistent.

The following two results are weaker:

Corollary 2. *Under the setup and assumptions made in Theorem 1, for n large,*

$$p^* \left\{ \pi(p : d(p^*, p) > 4\varepsilon_n | D^n) \geq 2e^{-n\varepsilon_n^2/4} \right\} \leq 2e^{-n\varepsilon_n^2/4},$$

$$\lim_{n \rightarrow \infty} p^* \left\{ \pi(p : d(p^*, p) \leq 4\varepsilon_n | D^n) \geq 1 - 2e^{-n\varepsilon_n^2/4} \right\} = 1$$

Proof. By Markov's inequality, the first result follows:

$$p^* \left\{ \pi(p : d(p^*, p) > 4\varepsilon_n | D^n) \geq 2e^{-n\varepsilon_n^2/4} \right\} \leq \frac{E^* \pi(d(p, p^*) > 4\varepsilon_n | D^n)}{2e^{-n\varepsilon_n^2/4}} \leq 2e^{-n\varepsilon_n^2/4}.$$

To show the second result, take complement of the first result:

$$p^* \left\{ \pi(p : d(p^*, p) > 4\varepsilon_n | D^n) < 2e^{-n\varepsilon_n^2/4} \right\} \geq 1 - 2e^{-n\varepsilon_n^2/4},$$

rewrite the set inside the $p^* \{ \}$,

$$p^* \left\{ \pi(p : d(p^*, p) \leq 4\varepsilon_n | D^n) \geq 1 - 2e^{-n\varepsilon_n^2/4} \right\} \geq 1 - 2e^{-n\varepsilon_n^2/4},$$

and then let $n \rightarrow \infty$, the second result follows. □

The Spike and Slab Prior and Convergence Results

3.1 Likelihood: Homoscedastic Gaussian Errors

The model considered in this chapter is the homoscedastic Gaussian error model which is an exponential family model considered in Chapter 2. The model is

$$\begin{aligned}x &= z^T \theta + \varepsilon_1 \\y &= \beta x + \varepsilon_2\end{aligned}$$

where $z \perp (\varepsilon_1, \varepsilon_2)$, $[\varepsilon_1, \varepsilon_2]^T \sim N(\mathbf{0}, \Sigma)$ and $\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix}$. The off-diagonal term σ_{12} is non-zero. Reparameterize the matrix Σ by Cholesky decomposition:

$$\Sigma = AHA^T$$

where $A = \begin{bmatrix} 1 & 0 \\ a_{21} & 1 \end{bmatrix}$, $H = \begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{2|1}^2 \end{bmatrix}$. The off-diagonal term of A , i.e. a_{21} , captures the strength of the endogeneity problem: the higher the $|a_{21}|$, the more inconsistency and bias from ordinary linear regression. (See Lopes and Polson (2014) for more discussion on this parameterization.)

3.2 The Spike and Slab Prior

With the likelihood function

$$f(x, y|\vartheta) \propto (\sigma_{11}^{-2} \sigma_{2|1}^{-2})^{1/2} \exp \left\{ -\frac{1}{2} \left\{ \sigma_{11}^{-2} (x - h)^2 + \sigma_{2|1}^{-2} (y - (a_{21} + \beta)x + a_{21}h)^2 \right\} \right\},$$

where $\vartheta = (h, \beta, a_{21}, \sigma_{11}^{-2}, \sigma_{2|1}^{-2})$, the Fisher information matrix of the likelihood,

$-E \left[\frac{\partial^2}{\partial \vartheta^2} \log f(x, y|\vartheta) \right]$, is near diagonal:

$$\begin{bmatrix} \sigma_{11}^{-2} + \sigma_{2|1}^{-2} a_{21}^2 & -h \sigma_{2|1}^{-2} a_{21} & 0 & 0 & 0 \\ -h \sigma_{2|1}^{-2} a_{21} & \sigma_{2|1}^{-2} (\sigma_{11}^2 + h^2) & \sigma_{2|1}^{-2} \sigma_{11}^2 & 0 & 0 \\ 0 & \sigma_{2|1}^{-2} \sigma_{11}^2 & \sigma_{2|1}^{-2} \sigma_{11}^2 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \sigma_{11}^4 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} \sigma_{2|1}^4 \end{bmatrix},$$

which indicates independent prior specification is desirable. The spike and slab prior specification is as follows:

Dependence Structure:

$$(\theta, \gamma) \perp\!\!\!\perp (\beta, \sigma_{2|1}^{-2}) \perp\!\!\!\perp \sigma_{11}^{-2} \perp\!\!\!\perp a_{21}$$

Prior on θ :

The prior on θ performs shrinkage and selection simultaneously: for $j = 1 : K_n$,

$$\theta_j | \gamma, \tau^2 \stackrel{iid}{\sim} \gamma_j N(0, \tau^2) + (1 - \gamma_j) 1_{\{\theta_j=0\}},$$

where the hyper prior on the parameter τ is $\tau^{-2} \sim TGa(s_\tau/2, S_\tau/2, t_\tau)$ which is a truncated Gamma distribution with support $[t_\tau, +\infty)$. The density of $W \sim TGa(a, b, t)$ is defined as

$$\pi(W = w) = \pi(X = w) 1_{\{X \geq t\}} / (1 - \pi(X < t)),$$

where $X \sim Ga(a, b)$ with support $(0, +\infty)$ and t is the truncation point. All the truncated Gamma distributions defined in this paper are truncated below, i.e., the support after truncation is $[t, +\infty)$.

The truncation is necessary in using Theorem 1. Without truncation, Normal-inverse Gamma conjugacy produces a t -distribution which has much heavier tail than sub-Gaussian tails. But we can truncate the inverse-Gamma distribution such that the induced marginal distribution has sub-Gaussian tails. The following prior specification follows this idea. As we can require that the truncation point to approach to 0, the truncation may be ignored in practice because sampling from a truncated Gamma distribution with arbitrarily small pieced truncated is essentially the same as the Gamma distribution with full support.

The prior on the model parameter γ is specified in the following.

Prior on model parameter γ :

The prior probability of model parameter γ being γ_n is

$$\pi(\gamma = \gamma_n) = \pi(|\gamma| = |\gamma_n|) / \binom{K_n}{|\gamma_n|}$$

where $\binom{k}{r}$ denotes the binomial coefficient, $|\gamma| \sim TPois - Gamma(\bar{r}_n, a_n, b_n)$ with support $\{0, 1, 2, \dots, \bar{r}_n\}$. $TPois - Gamma(r, a, b)$ denotes truncated Poisson-Gamma distribution with density

$$\pi(|\gamma| = k) = \int_0^\infty \frac{1}{\pi(X \leq r|\lambda)} \pi(X = k|\lambda) 1_{(k \leq r)} \pi(\lambda) d\lambda$$

where $X \sim Pois(\lambda)$ and $\lambda \sim Ga(a, b)$.

This prior can be generated in two steps. In the first step, generate the number of useful instruments over the support $\{0, 1, 2, \dots, \bar{r}_n\}$, say k , and in the second step, conditional on k , determine the location of the useful instruments: uniformly draw one model from the models of the same size k .

Prior on nuisance parameters ς :

The model has four nuisance parameters:

- The prior on $(\beta, \sigma_{2|1}^2)$ can be specified in two ways:
 - without conjugacy: $\beta \sim N(b_0, v_\beta)$, $\sigma_{2|1}^{-2} \sim TGa(s_2/2, S_2/2, t_{\sigma_2})$ and $\beta \perp\!\!\!\perp \sigma_{2|1}^{-2}$, where $t_{\sigma_2} \rightarrow 0$. The truncation is for proof's purpose, and in practice it affects little as t_{σ_2} is arbitrarily close to 0.
 - with conjugacy: $\beta | \sigma_{2|1}^2 \sim N(b_0, v_\beta \sigma_{2|1}^2)$, $\sigma_{2|1}^{-2} \sim TGa(s_2/2, S_2/2, t_\beta)$. t_β 's order is $1 < t_\beta^{-1} \lesssim n\varepsilon_n^2$. For instance, $t_\beta = 1/n\varepsilon_n^2$, t_β can be arbitrarily close to 0 as $n \rightarrow \infty$.
- $a_{21} \sim N(0, v_a^2)$ or $a_{21} | \sigma_{2|1}^{-2} \sim N(0, v_a^2 \sigma_{2|1}^2)$
- $\sigma_{11}^{-2} \sim TGa(s_1/2, S_1/2, t_{\sigma_1})$, where $t_{\sigma_1} \rightarrow 0$. The truncation is for proof's purpose, and in practice it affects little as t_{σ_1} is arbitrarily close to 0.

3.3 Posterior Consistency under Spike and Slab Priors

The posterior consistency is also in the sense of Hellinger distance. We further make the following assumptions on growth rates of K_n and \bar{r}_n :

Assumption R: Growth Rate Assumptions

(R1) $\bar{r}_n \ln(1/\varepsilon_n^2) < n\varepsilon_n^2$

(R2) $\bar{r}_n \ln K_n < n\varepsilon_n^2$

(R3) $\bar{r}_n^2/t_\tau \lesssim n\varepsilon_n^2$

(R4) $\bar{r}_n \ln D(\bar{r}_n, C_n) < n\varepsilon_n^2$, where $C_n \asymp n\varepsilon_n^2$.

The assumptions can be used in multiple ways. One could first specify the rate of ε_n , then use (R1) and (R4) to pin down the upper bound of the rate of \bar{r}_n , then use (R2) to pin down the upper bound of the rate of K_n , then use (R3) to pin down

the upper bound of the rate of t_τ . One could also first specify \bar{r}_n and K_n , then use (R1) to (R4) to pin down the lower bound of the rate of ε_n . Here K_n is allowed to grow exponentially with some polynomial of n so it can be much larger than n . But the growth rate affects the convergence rate negatively, which is characterized in Corollary 5. The cutoff t_τ is going to 0 and the rate is also characterized in Corollary 5.

Assumption S: Sparsity Assumption

(S1) The true coefficients on z are absolutely summable: $\lim_{n \rightarrow \infty} \sum_{j=1}^{K_n} |\theta_j^*| < \infty$, where K_n is nondecreasing in n .

(S2) $\Delta(r_n) < \varepsilon_n^2$ for any sequence $r_n > 1$, where $\Delta(r_n) = \inf_{\gamma: |\gamma|=r_n} \sum_{j: j \notin \gamma} |\theta_j^*|$.

Theorem 3 (Convergence rate under spike and slab prior). *The covariates of the model are bounded $|z_j| \leq 1$ for all $j = 1 : K_n$. Assumption R and Assumption S are made.*

If the prior is specified as in Section 3.2, then for n large enough,

$$E^* \pi(d(p, p^*) > 4\varepsilon_n | D^n) \leq 4e^{-n\varepsilon_n^2/2}.$$

Proof. Due to Theorem 1, it suffices to check Condition O and Condition N. The proof here is sketchy and the details of the calculations are in the Appendix A.2.

Checking Condition O(a)

Assertions $\bar{r}_n \ln(1/\varepsilon_n^2) < n\varepsilon_n^2$ and $\bar{r}_n \ln K_n < n\varepsilon_n^2$ are assumed to be valid.

It is left to show

$$\bar{r}_n \ln D(\bar{r}_n, C_n) < n\varepsilon_n^2.$$

To calculate $\ln D(\bar{r}_n, C_n)$, it is to calculate $A(h, \varsigma)$ and $\psi(h, \varsigma)$. Here $T(x, y)$ is chosen to be $[x^2, y^2, xy, x, y]^T$. The details of calculating $A(h, \varsigma)$ and $\psi(h, \varsigma)$ are in Appendix A.2.1.

The $A(h, \varsigma)$ matrix with respect to $(h, \beta, a_{21}, \sigma_{11}^{-2}, \sigma_{21}^{-2})$ is

$$\begin{bmatrix} 0 & 0 & 0 & \sigma_{11}^{-2} + \sigma_{21}^{-2} a_{21} (a_{21} + \beta) & -a_{21} \sigma_{21}^{-2} \\ -\sigma_{21}^{-2} (a_{21} + \beta) & 0 & \sigma_{21}^{-2} & h \sigma_{21}^{-2} a_{21} & 0 \\ -\sigma_{21}^{-2} (a_{21} + \beta) & 0 & \sigma_{21}^{-2} & h \sigma_{21}^{-2} (2a_{21} + \beta) & -h \sigma_{21}^{-2} \\ -1/2 & 0 & 0 & h & 0 \\ -(a_{21} + \beta)^2/2 & -1/2 & a_{21} + \beta & h a_{21} (a_{21} + \beta) & -h a_{21} \end{bmatrix},$$

whose elements have at most polynomial growth rate of C_n as $(h, \varsigma) \in B_{\bar{r}_n, C_n}^\infty$.

The vector $\psi(h, \varsigma)$ is

$$\psi(h, \varsigma) = [\sigma_{11}^2 + h^2, a_{21}^2 \sigma_{11}^2 + \sigma_{21}^2 + h^2 \beta^2, (a_{21} + \beta) \sigma_{11}^2 + h^2 \beta, h, h \beta]^T.$$

Since σ_{11}^{-2} and σ_{21}^{-2} are truncated below, there is no problem of 0 being denominator.

On the other extreme, elements in $\psi(h, \varsigma)$ have at most polynomial growth rate of C_n as $(h, \varsigma) \in B_{\bar{r}_n, C_n}^\infty$.

Because $C_n \asymp n \varepsilon_n^2$, it follows

$$\bar{r}_n \ln D(\bar{r}_n, C_n) < n \varepsilon_n^2.$$

Checking Condition O(b)

1. Checking $\pi(|\gamma| > \bar{r}_n) \leq e^{-4n\varepsilon_n^2}$

Notice $\pi(|\gamma| > \bar{r}_n) = 0$, this condition is trivially satisfied.

2. Checking $1 - \pi\left(\bigcap_{j=1:K_n} \{|\theta_j| < C_n\} | \gamma\right) \leq e^{-4n\varepsilon_n^2}$

Define $z_n = \max_{1 \leq j \leq r_n} |\theta_j|$ and the prior probability of at least one θ_j satisfying

$|\theta_j| > C_n$ becomes $\pi(z_n > C_n | \gamma)$:

$$\pi(z_n > C_n | \gamma) = 1 - \pi\left(\bigcap_{j=1:K_n} \{|\theta_j| < C_n\} | \gamma\right).$$

So the problem becomes verifying z_n has a sub-Gaussian tail.

Orlicz norm is a particularly useful tool here. Suppose X is some random variable and the Ψ -Orlicz norm is defined as

$$\|X\|_{\Psi} = \inf \{c > 0 : E\Psi(X/c) \leq 1\}$$

where $\Psi(\cdot)$ is some nondecreasing convex function. Orlicz norm theory provides bounds on tail probabilities:

$$\pi(|X| > C) \leq \Psi(C/\|X\|_{\Psi})^{-1},$$

provided the existence of Orlicz norms. (See Pollard (1990) for more details on Orlicz norm.)

Here let $\Psi(x) = e^{x^2}/5$, the inequality implied by Orlicz norm theory is

$$\pi(|z_n| > C_n) \leq 5e^{-C_n^2/\|z_n\|_{\Psi}^2},$$

provided the existence of the Orlicz norm. Suppose z_n 's Orlicz norm exists and $\|z_n\|_{\Psi}^2 < c$, c may be dependent on n or/and K_n , then

$$\pi(|z_n| > C_n) \leq 5e^{-C_n^2/\|z_n\|_{\Psi}^2} \leq 5e^{-C_n^2/c}.$$

Appendix A.2.2 shows that there exists c such that $C_n^2/c \gtrsim n\varepsilon_n^2$. The $\ln \bar{r}_n$ term is negligible because $1 < \bar{r}_n \leq K_n$ and $\ln \bar{r}_n \leq \bar{r}_n \ln K_n < n\varepsilon_n^2$. So the condition $1 - \pi\left(\bigcap_{j=1:K_n} \{|\theta_j| < C_n\}|\gamma\right) \leq \bar{r}_n e^{-4n\varepsilon_n^2}$ is satisfied.

3. Checking the nuisance parameters:

Checking $\pi(|a_{21}| > C_n) \leq e^{-4n\varepsilon_n^2}$

If independent prior is specified, we have a Normal prior. The Normal tail is the lightest and use Mill's ratio:

$$\pi(|a_{21}| > C_n) = 2(1 - \Phi(C_n/v_a)) \leq 2\phi(C_n/v_a) v_a/C_n = ce^{-\frac{1}{2v_a^2}C_n^2 - \ln C_n}$$

where c is a generic constant. Notice $C_n \asymp n\varepsilon_n^2$, the tail probability $\pi(|a_{21}| > C_n)$ is indeed bounded above by $e^{-4n\varepsilon_n^2}$ for n large enough.

If conjugate prior is specified, the proof is essentially the same as the conjugate case of β . So the details are omitted.

Checking $\pi(\sigma_{11}^{-2} > C_n) \leq e^{-4n\varepsilon_n^2}$

The truncated Gamma tail probability is heavier than Gamma tail without truncation but the inflation is bounded above by some finite number as the truncation point approaches 0. Notice

$$\pi(\sigma_{11}^{-2} > C_n) \leq c \int_{C_n}^{\infty} (\sigma_{11}^{-2})^{s_1/2-1} e^{-s_1\sigma_{11}^{-2}/2} \leq c \int_{C_n}^{\infty} (\sigma_{11}^{-2})^{s_1/2-1} e^{-\sigma_{11}^{-2}/2}$$

where c is a generic constant.

The χ^2 tail probability can be bounded, using the following result due to Lemma 1 of Laurent and Massart (2000):

For $X \sim \chi_k^2$ and k is the degree of freedom,

$$\pi(X - k \geq 2\sqrt{kx} + 2x) \leq e^{-x}.$$

Solve $C_n = s_1/2 + \sqrt{xs_1/2} + 2x$ for x and we have desired bound:

$$\pi(\sigma_{11}^{-2} > C_n) \leq ce^{-\frac{1}{4}(\sqrt{C_n - s_1/2} + \sqrt{s_1/2})^2} \asymp ce^{-\frac{1}{4}C_n},$$

where c is a generic constant.

So $C_n \asymp n\varepsilon_n^2$ implies the tail probability $\pi(\sigma_{11}^{-2} > C_n)$ is indeed bounded above by $e^{-4n\varepsilon_n^2}$ for n large enough.

Checking $\pi(|\beta| > C_n) \leq e^{-4n\varepsilon_n^2}$

If independence is specified on $(\beta, \sigma_{2|1}^{-2})$, the problem is checking a Normal tail. The proof is essentially the same as checking the tail probability of a_{21} . So the details are omitted.

If conjugate prior is assumed, the support of $\sigma_{2|1}^{-2}$ has to be truncated. This is because if the conjugacy is assumed for full support of $\sigma_{2|1}^{-2}$, the induced t -distribution tail is much heavier than sub-Gaussian tail.

Here let $\Psi(x) = e^{x^2}/5$, the inequality implied by Orlicz norm theory is

$$\pi(|\beta| > C_n) \leq 5e^{-C_n^2/\|\beta\|_\Psi^2},$$

provided the existence of the Orlicz norm.

Suppose $\|\beta\|_\Psi^2 < c^2$, c may be dependent on b_0 and v_β , then

$$\pi(|\beta| > C_n) \leq 5e^{-C_n^2/\|\beta\|_\Psi^2} \leq 5e^{-C_n^2/c^2}.$$

The condition $\pi(|\beta| > C_n) \leq e^{-4n\varepsilon_n^2}$ essentially requires $C_n^2/c^2 \gtrsim n\varepsilon_n^2$, so notice $C_n \asymp n\varepsilon_n^2$, it boils down to $c^2 \lesssim n\varepsilon_n^2$.

Appendix A.2.3 shows there exists c , such that $c^2 \lesssim n\varepsilon_n^2$ and the Orlicz-norm $\|\beta\|_\Psi^2 < c^2$. So the tail probability $\pi(|\beta| > C_n)$ is indeed bounded above by $e^{-4n\varepsilon_n^2}$ for n large enough.

Checking $\pi\left(\sigma_{2|1}^{-2} > C_n\right) \leq e^{-4n\varepsilon_n^2}$

If independence is specified on $(\beta, \sigma_{2|1}^{-2})$, the problem is checking a Gamma tail.

The proof is essentially the same as checking the tail probability of σ_{11}^{-2} . So the details are omitted.

If conjugate prior is assumed, the support of $\sigma_{2|1}^{-2}$ has to be truncated. As the truncation point is away from C_n , the tail probability of a truncated Gamma

distribution is proportional to the full support one:

$$\pi\left(\sigma_{2|1}^{-2} > C_n\right) = \frac{1}{1 - \pi(W \leq t_\beta)} \pi(W > C_n),$$

where $W \sim Ga(s_2/2, S_2/2)$.

It suffices to check the order of inflation factor $\frac{1}{1 - \pi(W \leq t_\beta)}$.

By monotonicity and continuity of $\pi(W \leq t_\beta)$ as a function of t_β , there exists $\underline{t} > 0$, such that $\forall t_\beta < \underline{t}$, $\pi(W \leq t_\beta) \leq \pi(W \leq \underline{t}) = \frac{1}{2}$.

So $\forall t_\beta < \underline{t}$, $\frac{1}{1 - \pi(W \leq t_\beta)} \leq 2$ and $\pi\left(\sigma_{2|1}^{-2} > C_n\right) \leq 2\pi(W > C_n)$. Notice $1 < t_\beta^{-1} \lesssim n\varepsilon_n^2$, the inflation factor is bounded above by some constant for n large enough.

Similar to σ_{11}^{-2} , $\pi(W > C_n) \leq ce^{-\frac{1}{4}(\sqrt{C_n - s_2/2} + \sqrt{s_2/2})^2} = e^{-\frac{1}{4}C_n}$, where c is some generic constant.

So $C_n \asymp n\varepsilon_n^2$ implies that for n large enough,

$$\pi\left(\sigma_{2|1}^{-2} > C_n\right) \leq e^{-4n\varepsilon_n^2}.$$

Checking Condition N(a)

Define a sequence of integers: $\{r_n\}_{n=1}^\infty$ such that $1 < r_n < K_n$ and $r_n \leq \bar{r}_n$. Define the sequence of models $\{\gamma_n\}_{n=1}^\infty$ such that $\sum_{j:j \notin \gamma_n} |\theta_j^*| = \inf_{\gamma:|\gamma|=r_n} \sum_{j:j \notin \gamma} |\theta_j^*|$. For ties, choose the model parameter with smallest sum of indices.

Notice the assumption $\Delta(r_n) < \varepsilon_n^2$, we have $\sum_{j \notin \gamma_n} |\theta_j^*| < \varepsilon_n^2$.

Checking Condition N(b)

1. Checking $\pi(\gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$.

To show $\pi(\gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$ for n large enough, it is equivalent to show $\ln \pi(\gamma = \gamma_n) \gtrsim -n\varepsilon_n^2$. The details are in Appendix A.2.4.

2. Checking $\pi((\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta} | \gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$.

Denote $M_{\gamma_n, \eta} = M_{\gamma_n, \eta}^\theta \cup M_{\gamma_n, \eta}^\varsigma$, $M_{\gamma_n, \eta}^\theta = (\theta_j^* \pm \eta\varepsilon_n^2/|\gamma_n|)_{j \in \gamma_n}$ and $M_{\gamma_n, \eta}^\varsigma = (\varsigma_j^* \pm \eta\varepsilon_n^2)_{j=1}^m$.

Notice the prior dependence structure $\theta \perp\!\!\!\perp \varsigma$, to show $\pi((\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta} | \gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$, it suffices to show

$$\ln \pi(\theta_\gamma \in M_{\gamma_n, \eta}^\theta | \gamma = \gamma_n) \gtrsim -n\varepsilon_n^2 \text{ and } \ln \pi(\varsigma \in M_{\gamma_n, \eta}^\varsigma | \gamma = \gamma_n) \gtrsim -n\varepsilon_n^2.$$

The the details of the calculus are in Appendix A.2.5.

□

Similar to the general theorem, the following two weaker results can be obtained.

Corollary 4. *Under the setup and assumptions made in Theorem 3, for n large,*

$$p^* \left\{ \pi(p : d(p^*, p) > 4\varepsilon_n | D^n) \geq 2e^{-n\varepsilon_n^2/4} \right\} \leq 2e^{-n\varepsilon_n^2/4},$$

$$\lim_{n \rightarrow \infty} p^* \left\{ \pi(p : d(p^*, p) \leq 4\varepsilon_n | D^n) \geq 1 - 2e^{-n\varepsilon_n^2/4} \right\} = 1$$

Proof. The proof is the same as Corollary 2, hence details are omitted.

□

The convergence rate depends on how fast K_n and \bar{r}_n grow. The infiniteness of K_n complicates the posterior convergence so as K_n grows very fast, the convergence rate is expected to be slow. \bar{r}_n is the maximum model size, describing how complicated the model can be. Larger \bar{r}_n indicates a larger model space to explore and intuitively, it needs more data to reach the same convergence rate.

Corollary 5 characterizes the convergence rate:

Corollary 5. *Under the same setup as in Theorem 3 and further assume for some $\delta_k, \delta_r > 0$, $\xi_k \geq \xi_r \geq 0$ and $\xi_k + \xi_r \in [0, 1)$ with some $C, C' > 0$,*

- $K_n \lesssim n^{\delta_k} \exp\{Cn^{\xi_k}\}$,
- $C' \ln n \leq \bar{r}_n < (\ln n)^{\delta_r} n^{\xi_r}$.

Then, we can take the convergence rate in Theorem 3 as

$$\varepsilon_n = n^{-\frac{1}{2}(1-\xi_k-\xi_r)} (\ln n)^{\frac{1}{2}(1+\delta_r)}.$$

Proof. By assumptions further made,

$$\bar{r}_n \ln K_n < n^{\xi_k+\xi_r} (\ln n)^{\delta_r+1}.$$

First match the rate relation $\bar{r}_n \ln K_n < n\varepsilon_n^2$, and ε_n can be chosen such that $n\varepsilon_n^2 = n^{\xi_k+\xi_r} (\ln n)^{\delta_r+1}$. Hence,

$$\varepsilon_n = n^{-\frac{1}{2}(1-\xi_k-\xi_r)} (\ln n)^{\frac{1}{2}(1+\delta_r)}.$$

With this, it is left to verify other rate assumptions made in Assumption R.

The truncation point t_τ can be chosen such that $\bar{r}_n^2/t_\tau \lesssim n\varepsilon_n^2$ and $1/t_\tau \gtrsim 1$. Let $t_\tau = (\ln n)^{-\delta_t} n^{-\xi_t}$, then $\bar{r}_n^2/t_\tau = \bar{r}_n^2 (\ln n)^{\delta_t} n^{\xi_t} < n\varepsilon_n^2 = n^{\xi_k+\xi_r} (\ln n)^{\delta_r+1}$ requires

$$(\ln n)^{2\delta_r} n^{2\xi_r} \lesssim n^{\xi_k+\xi_r-\xi_t} (\ln n)^{\delta_r-\delta_t+1}.$$

So either $\xi_t < \xi_k - \xi_r$ or $\xi_t = \xi_k - \xi_r$ with $\delta_t \leq 1 - \delta_r$.

For $\xi_k - \xi_r > 0$, we can choose $\xi_t \in (0, \xi_k - \xi_r)$; for $\xi_k - \xi_r = 0$, we can choose $\xi_t = 0$ with $\delta_t \in [0, 1 - \delta_r]$.

Other rate assumptions in Assumption R can also be verified. The details are in Appendix A.3.

□

Suppose $K_n \asymp n^\alpha$ where α is any positive number, there exists ξ_k arbitrarily close to 0 such that $K_n \leq e^{Cn^{\xi_k}}$. \bar{r}_n is chosen such that $\bar{r}_n = (\ln n)^{\delta_r}$ for some $\delta_r > 0$.

Then, there exists $\xi > \xi_k$ arbitrarily close to 0 such that

$$\varepsilon_n \asymp n^{-\frac{1}{2}(1-\xi_k)} (\ln n)^{\frac{1}{2}(1+\delta_r)} < n^{-\frac{1}{2}+\xi},$$

which is arbitrarily close to the finite dimensional convergence rate. The fast rate is achieved at the cost of the slowly growing \bar{r}_n . We are looking at a very small subset of potentially useful instruments.

If \bar{r}_n is chosen to grow fast such that $\bar{r}_n = n^{\xi_r} (\ln n)^{\delta_r}$ for some $\xi_r < \xi_k$ and $\delta_r > 0$. Then, there exists η arbitrarily close to 0 such that

$$\varepsilon_n \asymp n^{-\frac{1}{2}(1-\xi_k-\xi_r)} (\ln n)^{\frac{1}{2}(1+\delta_r)} < n^{-\frac{1}{2}+\xi_k+\xi_r+\eta}.$$

Indeed, the convergence rate is slower.

3.4 Finite Sample Posterior Inference

The above section offers theoretical guarantee for posterior consistency and convergence rate of a class of spike and slab priors. This section derives a Gibbs sampler to draw posterior samples from the posterior distribution.

3.4.1 Likelihood

The likelihood follows Section 3.1, that is, for $i = 1 : n$,

$$\begin{aligned} x_i &= z_i^T \theta + \varepsilon_{1i} \\ y_i &= \beta x_i + \varepsilon_{2i} \end{aligned}$$

where $\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \sim (\mathbf{0}, AHA^T)$, $A = \begin{bmatrix} 1 & 0 \\ a_{21} & 1 \end{bmatrix}$, $H = \begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{2|1}^2 \end{bmatrix}$.

Intercepts are omitted for simplicity. If intercepts are specified explicitly, improper priors are typically put on the intercepts and we can integrate out the intercepts, producing a demeaned version of the model.

The full data likelihood is

$$L(h, \varsigma) = \prod_{i=1}^n f(x_i, y_i, z_i | \theta, \gamma, \varsigma) \propto e^{\sum_{i=1}^n a(h_i, \varsigma) \cdot T(x_i, y_i) + b(h_i, \varsigma)},$$

where $h_i = z_i^T \theta$, $\sum_{i=1}^n b(h_i, \varsigma) = -\frac{1}{2} \left(\sigma_{11}^{-2} + \sigma_{2|1}^{-2} a_{21}^2 \right) \sum_{i=1}^n h_i^2 + \frac{n}{2} \ln \sigma_{11}^{-2} + \frac{n}{2} \ln \sigma_{2|1}^{-2}$, and

$$\begin{aligned} \sum_{i=1}^n a(h_i, \varsigma) \cdot T(x_i, y_i) &= -\frac{1}{2} \left(\sigma_{11}^{-2} + \sigma_{2|1}^{-2} (a_{21} + \beta)^2 \right) \sum_{i=1}^n x_i^2 - \frac{1}{2} \sigma_{2|1}^{-2} \sum_{i=1}^n y_i^2 \\ &\quad + \sigma_{2|1}^{-2} (a_{21} + \beta) \sum_{i=1}^n x_i y_i \\ &\quad + \left(\sigma_{11}^{-2} + \sigma_{2|1}^{-2} a_{21} (a_{21} + \beta) \right) \sum_{i=1}^n h_i x_i - a_{21} \sigma_{2|1}^{-2} \sum_{i=1}^n h_i y_i \end{aligned}$$

3.4.2 Prior Specification

The prior specification follows Section 3.2. We choose conjugate priors such that a Gibbs sampler can be derived without Metropolis-Hastings steps. In sampling truncated Gamma distribution, inverse CDF method is convenient and fast.

- For $j = 1 : K_n$,

$$\theta_j | \gamma, \tau^2 \stackrel{iid}{\sim} \gamma_j N(0, \tau^2) + (1 - \gamma_j) 1_{(\theta_j=0)},$$

where $\tau^{-2} \sim TGa(s_\tau/2, S_\tau/2, t_\tau)$.

- The prior probability of model parameter γ being γ_n is

$$\pi(\gamma = \gamma_n) = \pi(|\gamma| = |\gamma_n|) / \binom{K_n}{|\gamma_n|}$$

where $\binom{k}{r}$ denotes the binomial coefficient, $|\gamma| \sim TPois - Gamma(\bar{r}_n, a_n, b_n)$.

- $\beta | \sigma_{2|1}^2 \sim N(0, v_\beta \sigma_{2|1}^2)$, $a_{21} | \sigma_{2|1}^{-2} \sim N(0, v_a^2 \sigma_{2|1}^2)$
- $\sigma_{2|1}^{-2} \sim TGa\left(\frac{s_2}{2}, \frac{S_2}{2}, t_\beta\right)$, $\sigma_{11}^{-2} \sim TGa\left(\frac{s_1}{2}, \frac{S_1}{2}, 0\right)$.

3.4.3 Posterior Distribution: The Gibbs Sampler

The Gibbs sampler is composed of a series of full conditionals and by sampling conditional distributions, samples from the joint distribution are obtained. The full conditionals are

- Sampling $\sigma_{11}^{-2} | \text{others}$:

$$\sigma_{11}^{-2} | \theta \sim Ga \left(\frac{s_{1n}}{2}, \frac{S_{1n}}{2} \right)$$

where $s_{1n} = s_1 + n$ and $S_{1n} = S_1 + \sum_{i=1}^n (x_i - h_i)^2$.

- Sampling $\sigma_{21}^{-2} | \text{others}$:

$$\sigma_{21}^{-2} | \theta, \beta, a_{21} \sim T Ga \left(\frac{s_{2n}}{2}, \frac{S_{2n}}{2}, t_\beta \right)$$

where $s_{2n} = s_2 + n$, $S_{2n} = S_2 + \sum_{i=1}^n (a_{21} h_i + y_i - (a_{21} + \beta) x_i)^2$.

- Sampling $a_{21} | \text{others}$:

$$a_{21} | \theta, \beta, \sigma_{21}^2 \sim N (e_{an}, v_{an}^2 \sigma_{21}^2)$$

where $e_{an} = v_{an}^2 \sum_{i=1}^n (x_i - h_i) (y_i - \beta x_i)$ and $v_{an}^{-2} = \sum_{i=1}^n (x_i - h_i)^2 + v_a^{-2}$.

- Sampling $\beta | \text{others}$:

$$\beta | \theta, a_{21}, \sigma_{21}^2 \sim N (b_n, v_{\beta n}^2 \sigma_{21}^2)$$

where $b_n = v_{\beta n}^2 (\sum_{i=1}^n x_i (y_i - a_{21} (x_i - h_i)) + v_\beta^{-1} b_0)$ and $v_{\beta n}^{-2} = \sum_{i=1}^n x_i^2 + v_\beta^{-1}$

- Sampling $\theta | \text{others}$: for $j = 1 : K_n$,

$$\theta_j | \theta_{-j}, \tau \sim \gamma_j^* N (E_j, V_j^2) + (1 - \gamma_j^*) 1_{(\theta_j=0)}$$

where $\gamma_j^* \sim Ber (p_j^*)$, $p_j^* = \frac{p_j \rho_j}{p_j \rho_j + 1 - p_j}$.

$p_j \equiv \frac{(r_j+a_n)1_{(r_j+1 \leq \bar{r}_n)}}{(b_n+1)\bar{r}_n - b_n r_j + a_n}$ is the prior probability of θ_j drawn from the slab,

$r_j \equiv \sum_{\ell \neq j} \gamma_\ell = \sum_{\ell \neq j} 1_{(\theta_\ell \neq 0)}$ is the number of instruments already in the model,

$\rho_j \equiv \frac{N(0|0, \tau^2)}{N(0|E_j, V_j)}$ is the ratio of Normal distribution densities evaluated at 0,

$V_j^{-1} \equiv \sum_{i=1}^n z_{ij}^2 \left(\sigma_{11}^{-2} + a_{21}^2 \sigma_{21}^{-2} \right) + \tau^{-2}$ is the precision of the slab,

$h_{ij} \equiv \sum_{\ell \neq j} z_{i\ell} \theta_\ell = h_i - z_{ij} \theta_j$, and

$E_j \equiv V_j \sum_{i=1}^n z_{ij} \left[a_{21} \sigma_{21}^{-2} (a_{21} (x_i - h_{ij}) - (y_i - \beta x_i)) + \sigma_{11}^{-2} (x_i - h_{ij}) \right]$

The derivations are in Appendix A.4.

- Sampling $\tau^{-2} | \text{others}$

$$\tau^{-2} | \theta \sim TGa(s_{\tau n}/2, S_{\tau n}/2, t_\tau),$$

where $s_{\tau n} = s_\tau + \sum_{j=1}^{K_n} \gamma_j$, and $S_{\tau n} = S_\tau + \sum_{j=1}^{K_n} \theta_j^2$.

3.4.4 Computation Issues of The Gibbs Sampler

The Gibbs sampler may be computationally expensive. An alternative is to develop EM algorithm. Ročková and George (2014) develops an EM variable selection algorithm based on spike and slab priors. When the spike part is also a Normal distribution but with a much smaller variance, the EM algorithm has analytic forms and the EM algorithm is shown to converge quickly to a posterior mode. But when the spike part is a point mass, the EM algorithm does not have analytical forms. So we can use a sequence of spike and slab priors with the variance of the spike part shrinking towards 0.

A common caution is when the instruments are dependent, posterior distribution may be multimodal and the Gibbs sampler has very bad mixing. In this case, Metropolis-Hastings algorithms may be helpful in jumping away from the current

model. Swapping strategies are also helpful. (Geyer, 1991; Geyer and Thompson, 1995)

4

Monte Carlo Experiments

This chapter first illustrates an example how to do inference using a Gibbs sampler under a spike and slab prior. The data generating process and experiment design are described in section 4.1, the illustration of convergence and making inference by Gibbs sampler are described in section 4.3,

4.1 Data Generating Process

Data is $D^n = \{(y_i, x_i, z_i)\}_{i=1}^n$, where y_i is the response variable, x_i is the endogenous variable, z_i is the K_n -dimensional instrument variable vector, that is, there are K_n instrumental variable candidates at hand.

The data is generated by the following model: for $i = 1 : n$,

$$\begin{aligned}x_i &= z_i^T \theta + \varepsilon_{1i} \\y_i &= \beta x_i + \varepsilon_{2i}\end{aligned}$$

with $\beta = 1$, $\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right)$, z_i 's are Normally distributed random vectors with mean zero and variance-covariance matrix Σ_z where $\text{cov}(z_{ij_1}, z_{ij_2}) =$

$r^{|j_1-j_2|}$ with $r = 0.5$ or $r = 0$. The truth of $\theta_j = 0.7^{j-1}$, so the sparsity assumption is satisfied.

4.2 Prior Specification

The prior specification follows Section 3.4. When sample size is moderate and the instruments are not weak, the prior can be flat. But when sample size is small or the instruments are weak, the prior should not be too flat.

The following parameters are chosen to be moderately informative: $s_\tau = S_\tau = 2$ and $t_\tau = t_\beta = 1/\sqrt{n}$; $\bar{r}_n = \lfloor n^{\frac{1}{3}} \rfloor$, where $\lfloor \cdot \rfloor$ denotes floor operator and $a_n = b_n = 5$; $v_\beta = 100, v_a = 10, s_1 = s_2 = S_1 = S_2 = 5$. When sample size is moderate and the instruments are not weak, the moderately informative prior imposes little impact on the posterior inference. When sample size is small or the instruments are weak, the moderately informative prior makes posterior distribution not too diffusive.

4.3 Simulation Results

4.3.1 Estimation of (β, a_{21}) as $n \rightarrow \infty$

In this section, to investigate the estimation of (a_{21}, β) as $n \rightarrow \infty$, IV are generated independently. In Figure 4.1(a), sample size is 100 while the number of IV is 500. Marginally the estimation is satisfactory: both 95% credible intervals contain the true values, though the truth is not near the posterior mode. In Figure 4.1(b), sample size is increased to 200 while the number of IV is increased to 600. This manipulation is done to mimic the situation where the number of observations is increasing with the number of IV increasing as well. The theory tells us that K_n should not grow too fast, otherwise the convergence is slow. Indeed, the experiment design here does not increase K_n is too fast. Notice K_n is allowed to grow exponentially with n , increasing by the same amount is indeed not fast. Further, in Figure 4.1(c), the sample size is increased to 300 with K_n being increased to 700. As $\bar{r}_n = \lfloor n^{\frac{1}{3}} \rfloor$, the \bar{r}_n 's are

4,5,and 6 respectively for the three cases. The message is that as $n \rightarrow \infty$, even if K_n is increasing as well, we may obtain consistency and the estimation accuracy is increasing. This example illustrates the consistency results with sample size and IV number increasing.

Figure 4.1 presents the scatter plots of 10^4 draws from the posterior distributions for β and a_{21} . True value of $(\beta, a_{21}) = (1, 0.8)$ falls nearer the mode as the sample size increases (at a relatively faster rate than IV does).

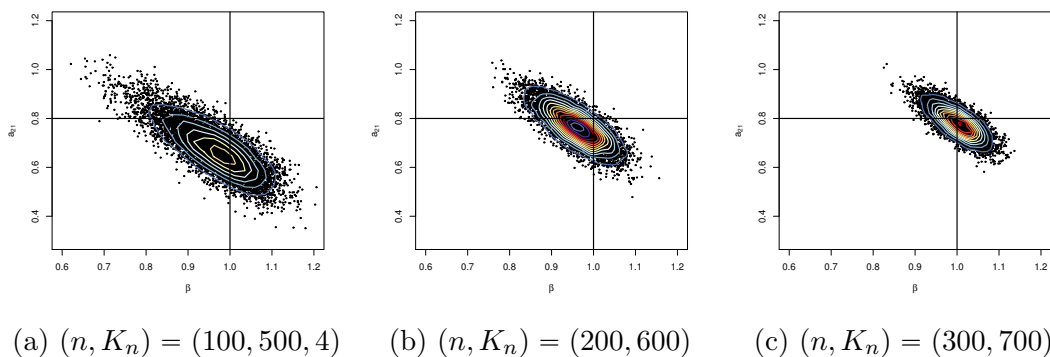


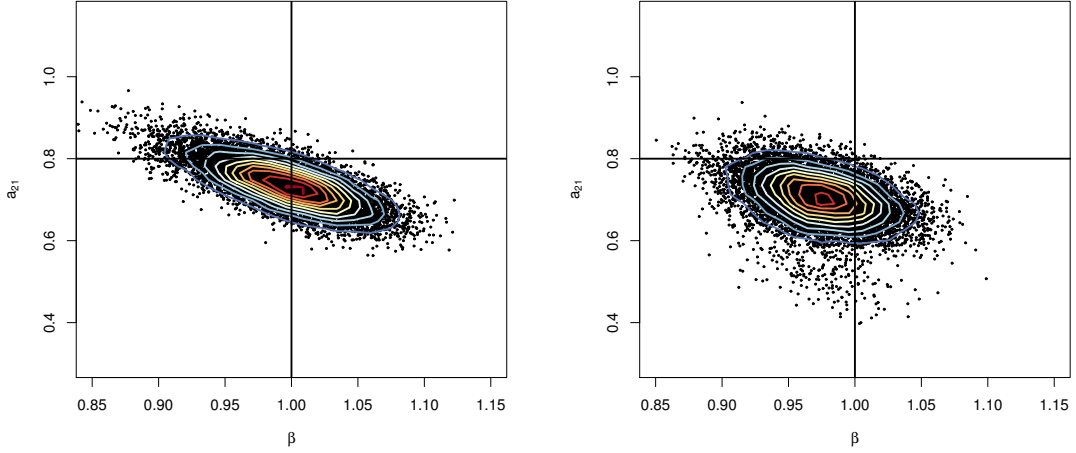
FIGURE 4.1: Scatter plots of (β, a_{21}) with independent IV

4.3.2 Estimation Under Independent/dependent IV

In practice, IV can be dependent. So to mimic positive dependence, different parameter values of r for variance-covariance matrix of Z are explored. The value of r measures the degree of correlation among instrumental variables and it is set to be 0.5 for the dependent case. In practice, the dependence can be very complicated so here is a very simple investigation. The parameter $r = 0.5$ imposes moderate dependence among IV at hand.

Figure 4.2(a) simulates another data with independent IV and scatters the samples from the Gibbs sampler. The plot has different axes from those of Figure 4.1(a). In Figure 4.2(b), the dependent case seems to have worse mixing and the

estimation for (β, a_{21}) seems worse. As the dependence increases, there will be more clouds of dots scattered because Gibbs sampler is slow in searching all the models and enumerating all the posterior modes.



(a) Independent IV

(b) Dependent IV

FIGURE 4.2: Scatter plots of (β, a_{21}) with $(n, K_n) = (300, 700)$

The presence of dependence makes the estimation of θ more complicated. Figure 4.3 plots the posterior mean $\hat{\theta}$ against true θ and the size of the circles represents the posterior inclusion probabilities. The posterior mean $\hat{\theta}$ is

$$\hat{\theta} = \frac{1}{N_{mc}} \sum_{t=1}^{N_{mc}} \theta^{(t)},$$

where $\theta^{(t)}$ is t^{th} iteration's draw of θ and the posterior inclusion probability is $\frac{1}{N_{mc}} \sum_{t=1}^{N_{mc}} \gamma^{(t)}$, where $\gamma^{(t)}$ is t^{th} iteration's draw of the model parameter.

The estimates $\hat{\theta}_j$, especially those of small θ_j 's, are shrunken to 0 more severely in the dependent case. Variable selection over a set of dependent variables remains an open question and the models explored in this paper also suffer from general critique of variable selection over dependent variables.

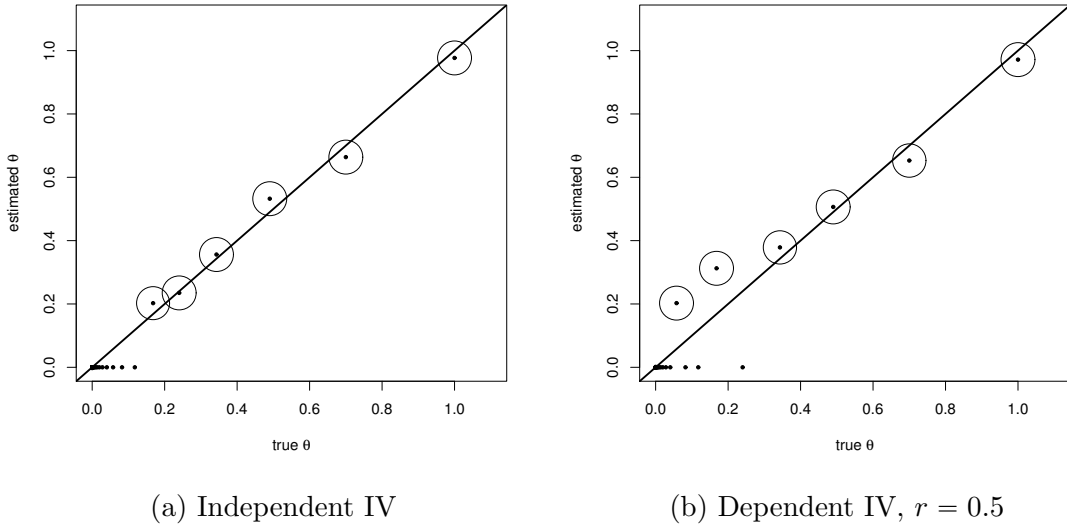


FIGURE 4.3: θ v.s. $\hat{\theta}$, $(n, K_n) = (300, 700)$

Figure 4.4 plots the posterior mean $\tilde{\theta}$ conditional on variables being chosen against true θ and the size of the circles represents the posterior inclusion probabilities.

$$\tilde{\theta}_j = \frac{1}{N_j} \sum_{t=1}^{N_{mc}} \theta_j^{(t)},$$

where $N_j = \sum_{t=1}^{N_{mc}} 1_{(\theta_j^{(t)} \neq 0)}$ and the subscript j denotes the j^{th} component of θ . $\tilde{\theta}_j$ is larger than $\hat{\theta}_j$ by construction, which is shown in Figure 4.4.

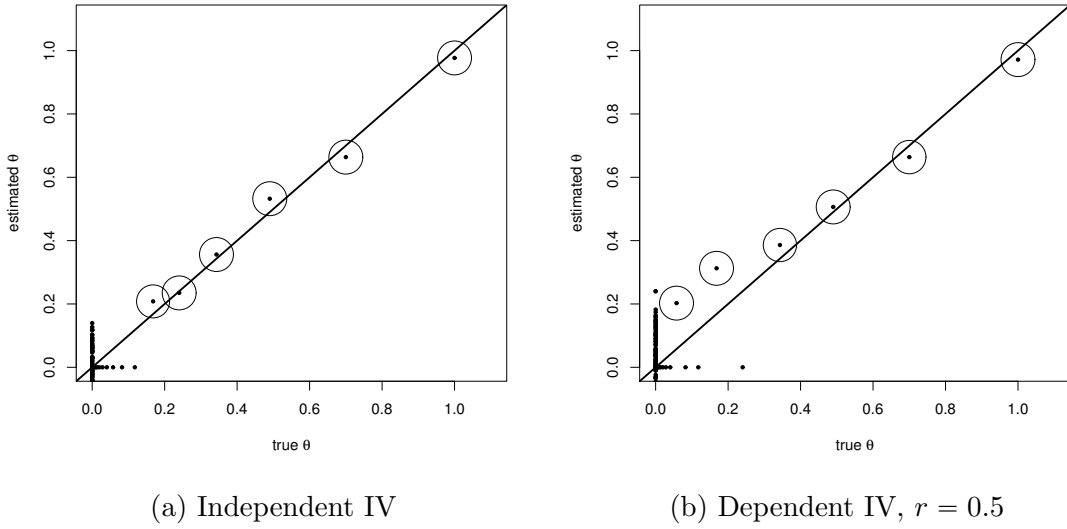


FIGURE 4.4: θ v.s. $\tilde{\theta}$, $(n, K_n) = (300, 700)$

The mixing of β and θ is examined in Appendix A.5. In the example, the mixing is generally good and robust to different starting values of $\beta^{(0)}$. Even though the beginning part is not similar to the later iterations for some plots, notice the burn-in is chosen to be 50, which is not large. So in practice, practitioners should check the mixing as well and decide how long the burn-in should be.

Discussion

This paper investigates Bayesian instrumental variable models with many potentially weak instruments. The number of instrumental variables grows with the sample size and is allowed to be much larger than the sample size. With some sparsity condition on the coefficients on the instruments, we characterize a general prior specification where the posterior consistency of the parameters is obtained and the corresponding convergence rate. In particular, we show the posterior consistency for the spike and slab prior on the many potentially weak instruments.

Future work may be comparing the Bayesian approach with other methods in the literature, e.g., TSLS/LIML/lasso based procedures and extending homoscedastic Gaussian error assumption to more flexible error assumptions. For instance, investigate the problem by Bayesian nonparametric approaches. The sparsity assumption $\lim_{n \rightarrow \infty} \sum_{j=1}^{K_n} |\theta_j^*| < \infty$ is crucial here. One future direction is to relax this assumption.

Appendix A

Appendix

A.1 Proof of Theorem 1

Before proving the theorem, we first introduce the crucial proposition:

Suppose \mathcal{P}_n is a sequence of sets of probability densities, ε_n is a sequence of positive numbers, $N(\varepsilon_n, \mathcal{P}_n)$ is the minimal number of Hellinger balls of radius ε_n to cover \mathcal{P}_n , data point $d_i \stackrel{iid}{\sim} p^*$ where p^* denotes the true density and its dimension $\dim(d_i)$ and p^* are allowed to depend on n . Denote $\pi(\cdot)$ as the prior density, denote $\pi(\cdot|D^n)$ as the posterior given data $D^n = (d_1, \dots, d_n)$.

Define $d_t(p^*, p) = t^{-1} (\int p^* (p^*/p)^t - 1)$ for $t > 0$. Denote $d_0(p_1, p_2)$ as the KL divergence.

Define the following conditions:

- (a) $\ln N(\varepsilon_n, \mathcal{P}_n) \leq n\varepsilon_n^2$ for n large.
- (b) $\pi(\mathcal{P}_n^c) \leq e^{-2n\varepsilon_n^2}$ for n large.
- (c) for $\gamma, r > 0$ small, $\exists N_{\gamma, r}$ such that $\forall n \geq N_{\gamma, r}$, $\pi(p : d_0(p^*, p) \leq \gamma\varepsilon_n^2) \geq e^{-rn\varepsilon_n^2}$.
- (d) $\exists t > 0$ such that $\pi(p : d_t(p^*, p) \leq \varepsilon_n^2/4) \geq e^{-n\varepsilon_n^2/4}$.

The proposition 1 in Jiang (2007) states that when $n\varepsilon_n^2 \rightarrow \infty$, condition (a),(b) and (d) imply that

$$E^* \pi (d(p^*, p) > 4\varepsilon_n |D^n) \leq 4e^{-n\varepsilon_n^2 \min\{1, t/2\}}$$

where the expectation is with respect to the true density p^* .

Now we proceed to the proof of Theorem 1.

Proof. It suffices to check the (a), (b) and (d) conditions in Proposition 1 with $t = 1$.

Checking Condition (a)

Denote \mathcal{P}_n as the set of densities with $|\gamma| \leq \bar{r}_n$ each $|\theta_j| < C_n$ and each $|\varsigma_j| < C_n$. The space \mathcal{P}_n can be covered by ℓ_∞ balls with radius δ of the form $B = (v_j \pm \delta)_{j=1}^{m+K_n}$. The radius and the corresponding number of such balls are shown below.

Consider the ℓ_∞ ball centering at $v = (v_j)_{j=1}^{m+|\gamma|}$ with radius δ . At most $\left(\frac{2C_n}{2\delta} + 1\right)^{m+|\gamma|}$ of such balls are needed to cover the parameter space of model γ . The number of models with size $|\gamma|$ in \mathcal{P}_n is $\binom{K_n}{|\gamma|} \leq K_n^{|\gamma|}$.

So to cover \mathcal{P}_n , $\sum_{|\gamma|=0}^{\bar{r}_n} K_n^{|\gamma|} \left(\frac{C_n}{\delta} + 1\right)^{m+|\gamma|}$ balls are sufficient and this implies $N(\delta)$ is bounded above by

$$(\bar{r}_n + 1) K_n^{\bar{r}_n} \left(\frac{C_n}{\delta} + 1\right)^{m+\bar{r}_n}.$$

Consider two densities p_u and p_v where p_u is in the ℓ_∞ ball centering at p_v with radius δ . Denote $p_w = f(y, x | \theta_w, \gamma_w, \varsigma_w) = e^{a(h_w, \varsigma_w) \cdot T(x, y) + b(h_w, \varsigma_w) + c(x, y)}$ and $I_w = a(h_w, \varsigma_w) \cdot \psi(h_w, \varsigma_w) + b(h_w, \varsigma_w)$ where $w = u, v$.

Then the KL divergence between p_u and p_v is

$$d_0(p_v, p_u) = E_z \int p_v \ln \frac{p_v}{p_u} \nu_y(dy) \nu_x(dx) = E_z (I_v - I_u).$$

By Taylor's theorem, a mean value expansion around the center p_v yields

$$\begin{aligned} I_u - I_v &= (a_1(\omega^i) \cdot \psi(h_v, \varsigma_v) + b_1(\omega^i))(h_u - h_v) \\ &\quad + \sum_{j=2}^{m+1} (a_j(\omega^i) \cdot \psi(h_v, \varsigma_v) + b_j(\omega^i)) (\zeta_u^{j-1} - \zeta_v^{j-1}) \end{aligned}$$

where $\omega^i = (h^i, \zeta^i)$ is an intermediate point between (h_v, ς_v) and (h_u, ς_u) .

Notice p_u is in the ℓ_∞ ball of p_v and $|z_j| \leq 1$ for all j ,

$$|h_v - h_u| = \left| \sum_{j \in \gamma} (\theta_{u,j} - \theta_{v,j}) z_j \right| \leq \sum_{j \in \gamma} |\theta_{u,j} - \theta_{v,j}| \leq \bar{r}_n \delta,$$

and $|\zeta_v^j - \zeta_u^j| \leq \delta$ for all $j = 1 : m$. These upper bounds are irrelevant to z .

Remember $a_j(h, \varsigma) \cdot \psi(h, \varsigma) + b_j(h, \varsigma) = 0$ for $j = 1 : J$.

Therefore,

$$d_0(p_v, p_u) \leq 2 \left[\max_{i,j} \sup_{(h,\varsigma) \in B_{\bar{r}_n, C_n}^\infty} |a_i^j(h, \varsigma)| \right] \left[\max_j \sup_{(h,\varsigma) \in B_{\bar{r}_n, C_n}^\infty} |\psi^j(h, \varsigma)| \right] (\bar{r}_n + m) \delta$$

Since Hellinger distance is bounded above by KL divergence:

$$d(p_v, p_u) \leq \sqrt{d_0(p_v, p_u)},$$

a uniform upper bound for the Hellinger distance between p_u and p_v is:

$$d(p_v, p_u) \leq \left\{ 2 \left[\max_{i,j} \sup_{(h,\varsigma) \in B_{\bar{r}_n, C_n}^\infty} |a_i^j(h, \varsigma)| \right] \left[\max_j \sup_{(h,\varsigma) \in B_{\bar{r}_n, C_n}^\infty} |\psi^j(h, \varsigma)| \right] (\bar{r}_n + m) \delta \right\}^{\frac{1}{2}}$$

So $d(p_v, p_u) \leq \varepsilon_n$, if

$$\delta = \varepsilon_n^2 / \left\{ 2 \left[\max_{i,j} \sup_{(h,\varsigma) \in B_{\bar{r}_n, C_n}^\infty} |a_i^j(h, \varsigma)| \right] \left[\max_j \sup_{(h,\varsigma) \in B_{\bar{r}_n, C_n}^\infty} |\psi^j(h, \varsigma)| \right] (\bar{r}_n + m) \right\}.$$

That is, density p_u in \mathcal{P}_n falls into the ε_n ball centered at p_v in the sense of Hellinger distance if p_u falls into the δ ball centered at p_v in the parameter space.

Notice there are at most $N(\delta)$ ball centers like p_v . The Hellinger covering number for \mathcal{P}_n with ball radius being ε_n is bounded above by $N(\delta)$, the covering number for the parameter space with ball radius being δ .

Remember we obtain a upper bound for $N(\delta)$ in the beginning, we have a upper bound for the covering number for the probability density space:

$$N(\varepsilon_n, \mathcal{P}_n) \leq N(\delta) \leq (\bar{r}_n + 1) K_n^{\bar{r}_n} \left(\frac{C_n}{\delta} + 1 \right)^{m + \bar{r}_n} \leq (2K_n^2 D(\bar{r}, C_n) / \varepsilon_n^2)^{m + \bar{r}_n}$$

With Condition O (a), for n large,

$$\ln N(\varepsilon_n, \mathcal{P}_n) < n\varepsilon_n^2$$

Checking Condition (b)

For the \mathcal{P}_n defined in checking (a), the prior probability mass put on its complement is bounded by

$$\begin{aligned} \pi(\mathcal{P}_n^c) &\leq \pi(|\gamma| > \bar{r}_n) + \pi\left(\bigcup_{j \in \gamma} \{|\theta_j| > C_n\} \mid |\gamma| \leq \bar{r}_n\right) \pi(|\gamma| \leq \bar{r}_n) \\ &\quad + \pi\left(\bigcup_{j=1:m} \{|\varsigma_j| > C_n\}\right) \\ &\leq \pi(|\gamma| > \bar{r}_n) + \bar{r}_n \cdot \max_{\gamma: |\gamma| \leq \bar{r}_n} \pi\left(\bigcup_{j \in \gamma} \{|\theta_j| > C_n\} \mid \gamma\right) \\ &\quad + m \cdot \max_{j=1:m} \pi(\{|\varsigma_j| > C_n\}) \end{aligned}$$

By Condition O, $\ln(m + \bar{r}_n + 1) \leq 2n\varepsilon_n^2$ for n large, because $\ln(\bar{r}_n + 1) \leq \bar{r}_n \ln K_n < n\varepsilon_n^2$ and m is a fixed number.

Then, condition (b) follows: for n large enough,

$$\pi(\mathcal{P}_n^c) \leq (m + \bar{r}_n + 1) e^{-4n\varepsilon_n^2} = e^{\ln(m + \bar{r}_n + 1) - 4n\varepsilon_n^2} \leq e^{-2n\varepsilon_n^2}$$

Checking Condition (d)

Take $t = 1$, for a generic model γ ,

$$d_t(p^*, p_\gamma) = \int p^* (p^*/p_\gamma) - 1 = \int p^* \left(e^{(a(h^*, \varsigma^*) - a(h_\gamma, \varsigma)) \cdot T(x, y) + b(h^*, \varsigma^*) - b(h_\gamma, \varsigma)} - 1 \right),$$

which is a integral over all (x, y, z) . Denote the integral

$$I(h_\gamma, \varsigma) = \int \int p_{x,y|z}^* \left(e^{(a(h^*, \varsigma^*) - a(h_\gamma, \varsigma)) \cdot T(x,y) + b(h^*, \varsigma^*) - b(h_\gamma, \varsigma)} - 1 \right) \nu_y(dy) \nu_x(dx),$$

which is a function of z and parameters (h_γ, ς) and is C^1 . We now establish a uniform upper bound for $I(h_\gamma, \varsigma)$. So a mean value expansion around (h^*, ς^*) yields

$$I(h_\gamma, \varsigma) = g_1(h^i, \varsigma^i)(h_\gamma - h^*) + \sum_{j=1}^m g_{j+1}(h^i, \varsigma^i)(\varsigma_j - \varsigma_j^*)$$

where $g_i(\cdot, \cdot)$ is the continuous partial derivative function with respect to the i^{th} component at the neighborhood of (h^*, ς^*) and (h^i, ς^i) is a intermediate point between (h_γ, ς) and (h^*, ς^*) .

Under Condition N, a sequence of models γ_n exists such that $\sum_{j \notin \gamma_n} |\theta_j^*| < \varepsilon_n^2$ as $n \rightarrow \infty$ and now we consider this sequence of model.

Suppose $(\theta_{\gamma_n}, \varsigma) \in M_{\gamma_n, \eta}$, notice $|z_j| \leq 1$ for all j , then for some small $\eta > 0$,

$$|h^i - h^*| \leq |h_{\gamma_n} - h^*| \leq \sum_{j \in \gamma_n} |z_j| |\theta_j - \theta_j^*| + \sum_{j \notin \gamma_n} |z_j| |\theta_j^*| \leq \eta \varepsilon_n^2 + \Delta_n,$$

where $\Delta_n = \sum_{j \notin \gamma_n} |\theta_j^*|$. So when n becomes large, $|h^i - h^*|$ is arbitrarily small. Also, $|h^i|$ is bounded as $|h^i| \leq |h^i - h^*| + |h^*|$ and $|h^*|$ is bounded due to $\lim_{n \rightarrow \infty} \sum_{j=1}^{K_n} |\theta_j^*| < \infty$.

Similarly, $|\varsigma_j^i - \varsigma_j^*| \leq \eta \varepsilon_n^2$ under Condition N implies $|\varsigma_j^i|$ is also bounded as $|\varsigma_j^i| \leq |\varsigma_j - \varsigma_j^*| + |\varsigma_j^*|$.

Observe each $g_i(\cdot, \cdot)$ is continuous, $|g_i(\cdot, \cdot)|$ is bounded above and denote a common upper bound as M . Then,

$$I(h_{\gamma_n}, \varsigma) \leq M(m+1)\eta \varepsilon_n^2 + M\Delta_n,$$

which has is irrelevant to z and we can choose η small enough and let n become large enough such that $M\Delta_n$ is negligible to ε_n^2 and $I(h_{\gamma_n}, \varsigma) \leq \frac{1}{4}\varepsilon_n^2$.

Notice $d_t(p^*, p_\gamma) = \int p_z^* I(h_\gamma, \varsigma) \nu_z(dz)$ by definition, $d_t(p^*, p_{\gamma_n}) \leq \frac{1}{4}\varepsilon_n^2$, that is, all the densities p_γ with $(\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta}$ satisfy $d_t(p^*, p_\gamma) \leq \frac{1}{4}\varepsilon_n^2$.

In set notation, $\{p_\gamma : (\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta}, \gamma = \gamma_n\} \subseteq \{p : d_t(p^*, p) \leq \varepsilon_n^2/4\}$. Condition N assumes,

$$\pi((\theta_{\gamma_n}, \varsigma) \in M_{\gamma_n, \eta}) = \pi((\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta} | \gamma = \gamma_n) \pi(\gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/4},$$

which immediately implies $\pi(p : d_t(p^*, p) \leq \varepsilon_n^2/4) \geq e^{-n\varepsilon_n^2/4}$.

□

A.2 Proof of Theorem 3

A.2.1 Likelihood Calculation

The likelihood is

$$\begin{aligned} f(x, y | h, \varsigma) &= (2\pi)^{-1} |\tilde{\Sigma}|^{-1/2} e^{-\frac{1}{2} \begin{bmatrix} x-h \\ y-h, \beta \end{bmatrix}^T \tilde{\Sigma}^{-1} \begin{bmatrix} x-h \\ y-h, \beta \end{bmatrix}} \\ &= (2\pi)^{-1} (\sigma_{11}^{-2} \sigma_{21}^{-2})^{1/2} e^{-\frac{1}{2} \{ \sigma_{11}^{-2} (x-h)^2 + \sigma_{21}^{-2} (y - (a_{21} + \beta)x + a_{21}h)^2 \}} \end{aligned}$$

Let $T(x, y) = [x^2, y^2, xy, x, y]^T$, then

$$a(h, \beta, a_{21}, \sigma_{11}^{-2}, \sigma_{21}^{-2}) = \begin{bmatrix} -\frac{1}{2} (\sigma_{11}^{-2} + \sigma_{21}^{-2} (a_{21} + \beta)^2) \\ -\frac{1}{2} \sigma_{21}^{-2} \\ \sigma_{21}^{-2} (a_{21} + \beta) \\ h (\sigma_{11}^{-2} + \sigma_{21}^{-2} a_{21} (a_{21} + \beta)) \\ -h a_{21} \sigma_{21}^{-2} \end{bmatrix},$$

and $b(h, \beta, a_{21}, \sigma_{11}^{-2}, \sigma_{21}^{-2}) = -\frac{h^2}{2} (\sigma_{11}^{-2} + \sigma_{21}^{-2} a_{21}^2) + \frac{1}{2} \ln \sigma_{11}^{-2} + \frac{1}{2} \ln \sigma_{21}^{-2}$.

The $A(h, \varsigma)$ matrix with respect to $(h, \beta, a_{21}, \sigma_{11}^{-2}, \sigma_{21}^{-2})$ is

$$\begin{bmatrix} 0 & 0 & 0 & \sigma_{11}^{-2} + \sigma_{21}^{-2} a_{21} (a_{21} + \beta) & -a_{21} \sigma_{21}^{-2} \\ -\sigma_{21}^{-2} (a_{21} + \beta) & 0 & \sigma_{21}^{-2} & h \sigma_{21}^{-2} a_{21} & 0 \\ -\sigma_{21}^{-2} (a_{21} + \beta) & 0 & \sigma_{21}^{-2} & h \sigma_{21}^{-2} (2a_{21} + \beta) & -h \sigma_{21}^{-2} \\ -1/2 & 0 & 0 & h & 0 \\ -(a_{21} + \beta)^2/2 & -1/2 & a_{21} + \beta & h a_{21} (a_{21} + \beta) & -h a_{21} \end{bmatrix}.$$

The matrix $B(h, \varsigma)$ with respect to $(h, \beta, a_{21}, \sigma_{11}^{-2}, \sigma_{2|1}^{-2})$ is

$$B(h, \varsigma) = \left[-h \left(\sigma_{11}^{-2} + \sigma_{2|1}^{-2} a_{21}^2 \right), 0, -h^2 \sigma_{2|1}^{-2} a_{21}, \frac{1/\sigma_{11}^{-2} - h^2}{2}, \frac{1/\sigma_{2|1}^{-2} - h^2 a_{21}^2}{2} \right]^T.$$

With bivariate Normality, we do not need to solve the linear system:

$$A(h, \varsigma) \psi(h, \varsigma) + B(h, \varsigma) = 0.$$

The expectation of $T(x, y)$ is

$$\begin{aligned} \psi(h, \varsigma) &= E_{(h, \varsigma)} T(x, y) \\ &= \left[\sigma_{11}^2 + h^2, a_{21}^2 \sigma_{11}^2 + \sigma_{2|1}^2 + h^2 \beta^2, (a_{21} + \beta) \sigma_{11}^2 + h^2 \beta, h, h\beta \right]^T. \end{aligned}$$

The invertibility of $A(h, \varsigma)$ requires $h \neq 0$, that is, at least one instrument is useful, or at least one coefficient on the instruments is not zero.

A.2.2 Checking $\pi(|\gamma| > \bar{r}_n) \leq e^{-4n\varepsilon_n^2}$

The following proof shows there exist c , such that $C_n^2/c \gtrsim n\varepsilon_n^2$ and the Orlicz-norm $\|z_n\|_{\Psi}^2 < c$. First verify the existence of the the Orlicz-norm $\|z_n\|_{\Psi}^2$ treating C_n as fixed by showing an upper bound and then invoke $C_n \asymp n\varepsilon_n^2$ to obtain the correct rate.

First notice $\pi(z_n \leq t|\tau, \gamma_n) = \pi(|\theta_j| \leq t|\tau, \gamma_n)^{r_n} = (2\Phi(t/\tau) - 1)^{r_n}$, where $\Phi(\cdot)$ denotes c.d.f. of standard Normal distribution. Then the probability density function of z_n conditional on τ and the model γ_n is

$$\pi(z_n = t|\tau, \gamma_n) = 2r_n(2\Phi(t/\tau) - 1)^{r_n-1} \phi(t/\tau)/\tau,$$

where $r_n = |\gamma_n| = \sum_j \gamma_{n,j}$ is the number of valid instruments selected in the model and $\phi(\cdot)$ denotes p.d.f. of standard Normal distribution.

Then we verify the existence of the Ψ -Orlicz norm conditional on τ .

$$\begin{aligned} E(\Psi(z_n/c) | \tau) &= \int_0^\infty \Psi(x/c) \pi(z_n = x) dx \\ &= \int_0^{r_n \tau} \Psi(x/c) \pi(z_n = x) dx + \int_{r_n \tau}^\infty \Psi(x/c) \pi(z_n = x) dx, \end{aligned}$$

where the truncation $r_n \tau$ can be replaced by $r_n^a \tau$ with $a \in (0, 1)$, e.g., $\sqrt{r_n} \tau$, to get a slightly sharper bound.

Notice $\Psi(\cdot)$ is increasing, $\pi(z_n \leq r_n \tau) \leq 1$ and $(2\Phi(\cdot) - 1)^{r_n^{-1}} \leq 1$,

$$E(\Psi(z_n/c) | \tau) \leq \frac{1}{5} e^{(r_n \tau/c)^2} + \int_{r_n \tau}^\infty \frac{1}{5} e^{(x/c)^2} \frac{2r_n}{\tau} \phi(x/\tau) dx.$$

Notice the last integral is

$$\begin{aligned} \int_{r_n \tau}^\infty \frac{1}{5} e^{(x/c)^2} \frac{2r_n}{\tau} \phi\left(\frac{x}{\tau}\right) dx &= \frac{2r_n}{5\tau} \int_{r_n \tau}^\infty e^{(x/c)^2} \frac{1}{\sqrt{2\pi}} e^{-(x/\tau)^2/2} dx \\ &= \frac{2r_n}{5\tau} \int_{r_n \tau}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2(\tau^{-2} - 2c^{-2})} dx \\ &= \frac{2r_n}{5\tau} \frac{1}{\sqrt{\tau^{-2} - 2c^{-2}}} \left(1 - \Phi\left(r_n \tau \sqrt{\tau^{-2} - 2c^{-2}}\right)\right) \\ &= \frac{2r_n}{5\sqrt{1 - 2\tau^2/c^2}} \left(1 - \Phi\left(r_n \sqrt{1 - 2\tau^2/c^2}\right)\right) \end{aligned}$$

Further by Mill's ratio,

$$\int_{r_n \tau}^\infty \frac{1}{5} e^{(x/c)^2} \frac{2r_n}{\tau} \phi\left(\frac{x}{\tau}\right) dx \leq \frac{2}{5(1 - 2\tau^2/c^2)} \phi\left(r_n \sqrt{1 - 2\tau^2/c^2}\right).$$

Therefore,

$$E\left(\Psi\left(\frac{z_n}{c}\right) | \tau\right) \leq \frac{1}{5} e^{(r_n \tau/c)^2} + \frac{2}{5(1 - 2\tau^2/c^2)} \phi\left(r_n \sqrt{1 - 2\tau^2/c^2}\right).$$

Let $c = \max(r_n, 5)\tau$, for example, $E\Psi(z_n/c) < 1$. So we obtain an upper bound for the Orlicz norm conditional on τ :

$$\|z_n\|_{\Psi|\tau} \leq \max(r_n, 5) \tau.$$

Hence, we obtain an upper bound for the tail probability:

$$\pi(|z_n| > C_n | \tau, \gamma) \leq 5e^{-C_n^2/\|z_n\|_{\Psi|\tau}^2} \leq 5e^{-C_n^2 \max(r_n, 5)^{-2} \tau^{-2}}.$$

The next step is to integrate out τ over $(t_\tau, +\infty)$. This truncation rules out the prior that the variances of the coefficients on the instruments are infinity. Intuitively, the prior belief of the truncation is that the the instruments' effects cannot vary too wildly.

Integrating out τ for some t_τ can produce the desired tail probability:

$$\pi(|z_n| > C_n | \gamma) = \int_{t_\tau}^{\infty} \pi(|z_n| > C_n | \tau, \gamma) \pi(\tau^{-2}) d\tau^{-2} \leq e^{-C_n^2 \max(\bar{r}_n, 5)^{-2} t_\tau + \ln 5}.$$

Notice $\bar{r}_n > 1$, the upper bound is essentially $e^{-C_n^2 t_\tau / \bar{r}_n^2 + \ln 5}$.

So with $C_n = n\varepsilon_n^2$ and $\bar{r}_n^2/t_\tau \lesssim n\varepsilon_n^2$, $C_n^2 t_\tau / \bar{r}_n^2 \gtrsim n\varepsilon_n^2$ and for n large enough,

$$1 - \pi\left(\bigcap_{j=1:K_n} \{|\theta_j| < C_n\} | \gamma\right) \leq e^{-4n\varepsilon_n^2}.$$

A.2.3 Checking $\pi(|\beta| > C_n) \leq e^{-4n\varepsilon_n^2}$

For $X \sim N(\mu, \sigma^2)$, we have

$$\begin{aligned} E\Psi(X/c) &= \int_{-\infty}^{\infty} \Psi(x/c) N(x|\mu, \sigma^2) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{5} e^{x^2/c^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \frac{1}{5} \frac{c}{\sqrt{c^2 - 2\sigma^2}} e^{\mu^2/(c^2 - 2\sigma^2)} \end{aligned}$$

So conditional on $\sigma_{2|1}^2$, $\beta|\sigma_{2|1}^2 \sim N(b_0, v_\beta \sigma_{2|1}^2)$ implies

$$E(\Psi(\beta/c) | \sigma_{2|1}^2) = \frac{1}{5} \frac{c}{\sqrt{c^2 - 2v_\beta \sigma_{2|1}^2}} e^{b_0^2/(c^2 - 2v_\beta \sigma_{2|1}^2)}.$$

Let $c^2 = b_0^2 + 5v_\beta \sigma_{2|1}^2$, then

$$E(\Psi(\beta/(b_0^2 + 5v_\beta \sigma_{2|1}^2)) | \sigma_{2|1}^2) \leq \frac{1}{5} \frac{1}{\sqrt{1 - 2/5}} e < 1.$$

So with the upper bound, for $\sigma_{2|1}^{-2} \geq t_\beta$,

$$\pi(|\beta| > C_n | \sigma_{2|1}^2) \leq 5e^{-C_n^2/c^2} = 5e^{-C_n^2/(b_0^2 + 5v_\beta \sigma_{2|1}^2)} \leq 5e^{-C_n^2/(b_0^2 + 5v_\beta t_\beta^{-1})}.$$

Then the tail probability

$$\pi(|\beta| > C_n) = \int_{t_\beta}^{\infty} \pi(|\beta| > C_n | \sigma_{2|1}^2) \pi(\sigma_{2|1}^{-2}) d\sigma_{2|1}^{-2} \leq e^{\ln 5 - C_n^2/(b_0^2 + 5v_\beta t_\beta^{-1})}$$

has desired rate: $C_n \asymp n\varepsilon_n^2$ and $t_\beta^{-1} \lesssim n\varepsilon_n^2$ implies $C_n^2/(b_0^2 + 5v_\beta t_\beta^{-1}) \gtrsim n\varepsilon_n^2$.

A.2.4 Checking $\pi(\gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$

First notice the Poisson-Gamma mixture is a negative binomial distribution, so the truncated Poisson-Gamma density has a lower bound of negative binomial density over the support with positive probability:

$$\begin{aligned} \pi(|\gamma_n| = r_n) &= \int_0^{\infty} \frac{1}{\pi(X \leq \bar{r}_n | \lambda)} \pi(X = r_n | \lambda) 1_{\{r_n \leq \bar{r}_n\}} \pi(\lambda) d\lambda / \binom{K_n}{r_n} \\ &\geq \int_0^{\infty} \pi(X = r_n | \lambda) 1_{\{r_n \leq \bar{r}_n\}} \pi(\lambda) d\lambda / \binom{K_n}{r_n} \\ &= \frac{\Gamma(a_n + r_n)}{\Gamma(a_n) r_n!} \left(1 - \frac{1}{b_n + 1}\right)^{a_n} \left(\frac{1}{b_n + 1}\right)^{r_n} / \binom{K_n}{r_n} \end{aligned}$$

Notice the assumption $\bar{r}_n \ln K_n < n\varepsilon_n^2$, $r_n < K_n$ and $r_n \leq \bar{r}_n$,

$$\begin{aligned} \ln \left(1 / \binom{K_n}{r_n}\right) &\sim r_n \ln(r_n / K_n) + (K_n - r_n) \ln(1 - r_n / K_n) \\ &\sim r_n \ln(r_n / K_n) + r_n \\ &\geq -\bar{r}_n \ln(K_n) > -n\varepsilon_n^2 \\ \ln \frac{\Gamma(a_n + r_n)}{(b_n + 1)^{r_n} r_n!} &\sim \ln \frac{(a_n + r_n)^{a_n + r_n}}{(b_n + 1)^{r_n} r_n!} = r_n \ln \left(1 + \frac{a_n}{r_n}\right) + a_n \ln(a + r_n) - r_n \ln(b_n + 1) \\ &\sim -r_n \ln(b_n + 1) \\ &\geq -\bar{r}_n \ln(K_n) > -n\varepsilon_n^2 \end{aligned}$$

So for n large enough, $\pi(\gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$

A.2.5 Checking $\pi((\theta_\gamma, \varsigma) \in M_{\gamma_n, \eta} | \gamma = \gamma_n) \geq e^{-n\varepsilon_n^2/8}$

Checking the parameters on instruments

Since $\theta_j | j \in \gamma_n, \tau \stackrel{iid}{\sim} N(0, \tau^2)$,

$$\pi(\theta_\gamma \in M_{\gamma_n, \eta}^\theta | \gamma = \gamma_n, \tau) \geq (2\eta\varepsilon_n^2/r_n)^{r_n} (2\pi)^{-r_n/2} (\tau^{-2})^{r_n/2} e^{-\frac{1}{2}\tau^{-2}\sum_{j \in \gamma_n} \hat{\theta}_j^2},$$

where $\hat{\theta}_{\gamma_n} \in \bar{M}_{\gamma_n, \eta}^\theta$ is bounded for n large enough and the upper bar over a set denotes its closure. Denote an upper bound as B_θ .

We next estimate an lower bound for

$$I_\theta \equiv \int_{t_\tau}^{\infty} (\tau^{-2})^{(r_n+s_\tau)/2-1} e^{-\frac{1}{2}\tau^{-2}(S_\tau+B_\theta)} d\tau^{-2}.$$

Notice for t near 0,

$$\int_0^t x^{a-1} e^{-bx} dx \approx t^a/a,$$

then since $t_\tau \rightarrow 0$,

$$\int_0^{t_\tau} (\tau^{-2})^{(r_n+s_\tau)/2-1} e^{-\frac{1}{2}\tau^{-2}(S_\tau+B_\theta)} d\tau^{-2} \approx 2t_\tau^{(r_n+s_\tau)/2} / (r_n + s_\tau).$$

Also notice the integral

$$\int_0^{\infty} (\tau^{-2})^{\frac{r_n+s_\tau}{2}-1} e^{-\frac{S_\tau+B_\theta}{2}\tau^{-2}} d\tau^{-2} = \Gamma\left(\frac{r_n + s_\tau}{2}\right) \left(\frac{S_\tau + B_\theta}{2}\right)^{-\frac{r_n+s_\tau}{2}},$$

which can be applied Stirling's approximation.

So for n large enough, r_n will be large enough and

$$\begin{aligned} I_\theta &= \int_0^{\infty} - \int_0^{t_\tau} (\tau^{-2})^{(r_n+s_\tau)/2-1} e^{-\frac{1}{2}\tau^{-2}(S_\tau+B_\theta)} d\tau^{-2} \\ &\approx \sqrt{\pi(r_n + s_\tau)} \left(\frac{r_n+s_\tau}{e(S_\tau+B_\theta)}\right)^{\frac{1}{2}(r_n+s_\tau)} - 2\frac{t_\tau^{(r_n+s_\tau)/2}}{r_n+s_\tau} \\ &\geq (e(S_\tau + B_\theta))^{-\frac{1}{2}(r_n+s_\tau)} \end{aligned}$$

Now we can check $\ln \pi(\theta_\gamma \in M_{\gamma_n, \eta}^\theta | \gamma = \gamma_n) \gtrsim -n\varepsilon_n^2$,

$$\begin{aligned} \pi(\theta_\gamma \in M_{\gamma_n, \eta}^\theta | \gamma = \gamma_n) &= \int_{t_\tau}^{\infty} \pi(\theta_\gamma \in M_{\gamma_n, \eta}^\theta | \gamma = \gamma_n, \tau) \pi(\tau^{-2}) d\tau^{-2} \\ &\geq (2\eta\varepsilon_n^2/r_n)^{r_n} (2\pi)^{-r_n/2} \int_{t_\tau}^{\infty} (\tau^{-2})^{r_n/2} e^{-\frac{B_\theta}{2}\tau^{-2}} \pi(\tau^{-2}) d\tau^{-2} \\ &\geq c(2\eta\varepsilon_n^2/r_n)^{r_n} (2\pi)^{-r_n/2} I_\theta, \end{aligned}$$

where c is a generic constant and the second inequality is due to truncation inflates the probability density where it has positive density after truncation.

So with $r_n \leq \bar{r}_n < K_n$, for n large enough,

$$\begin{aligned} \ln \pi (\theta_\gamma \in M_{\gamma_n, \eta}^\theta | \gamma = \gamma_n) &\geq c - r_n \ln r_n - r_n \ln (1/\varepsilon_n^2) \\ &\quad - r_n \left[\ln 2\eta + \frac{1}{2} \ln (2\pi e (S_\tau + B_\theta)) \right] \\ &\geq c - 2\bar{r}_n \ln K_n - \bar{r}_n \ln (1/\varepsilon_n^2), \end{aligned}$$

where c is a generic constant.

Notice $\bar{r}_n \ln K_n < n\varepsilon_n^2$ and $\bar{r}_n \ln (1/\varepsilon_n^2) < n\varepsilon_n^2$, so for n large enough,

$$\ln \pi (\theta_\gamma \in M_{\gamma_n, \eta}^\theta | \gamma = \gamma_n) > -n\varepsilon_n^2.$$

Checking the nuisance parameters

$\pi(\cdot)$ denotes the prior density function which is continuous,

$$\pi (\zeta \in M_{\gamma_n, \eta}^\zeta) = \int_{\zeta^* - \eta\varepsilon_n^2}^{\zeta^* + \eta\varepsilon_n^2} \pi (t) dt \geq (2\eta\varepsilon_n^2)^m \pi (\hat{\zeta}),$$

where $\hat{\zeta} \in \bar{M}_{\gamma_n, \eta}^\zeta$ and the upper bar over a set denotes its closure. So

$$\ln \pi (\zeta \in M_{\gamma_n, \eta}^\zeta) \geq c - m \ln (1/\varepsilon_n^2) + \ln \pi (\hat{\zeta})$$

Notice $\hat{\zeta} \in \bar{M}_{\gamma_n, \eta}^\zeta$ is bounded and the prior distribution for ζ is continuous, so $\ln \pi (\hat{\zeta})$ is bounded.

Notice the assumption $\bar{r}_n \ln (1/\varepsilon_n^2) < n\varepsilon_n^2$ and $1 < \bar{r}_n$, for n large enough,

$$m \ln (1/\varepsilon_n^2) \leq \bar{r}_n \ln (1/\varepsilon_n^2) < n\varepsilon_n^2.$$

A.3 Verify Other Rate Assumptions

- $\bar{r}_n \ln (1/\varepsilon_n^2) < n\varepsilon_n^2$:

$$\bar{r}_n \ln (1/\varepsilon_n^2) = \bar{r}_n \ln n < n^{\xi_k + \xi_r} (\ln n)^{\delta_r + 1} \asymp n\varepsilon_n^2.$$

- $\bar{r}_n \ln D(\bar{r}_n, C_n) < n\varepsilon_n^2$:

since $C_n \asymp n\varepsilon_n^2$ and $D(\bar{r}_n, C_n)$ is a polynomial of C_n ,

$$\bar{r}_n \ln D(\bar{r}_n, C_n) \asymp \bar{r}_n \ln(n\varepsilon_n^2) < \bar{r}_n \ln n < (\ln n)^{\delta_r+1} n^{\xi_r} \lesssim n\varepsilon_n^2.$$

A.4 Derivation of The Full Conditional of θ_j

First calculate the ratio:

$$\frac{p_j}{1-p_j} = \frac{\pi(\gamma_j = 1|\gamma_{-j})}{\pi(\gamma_j = 0|\gamma_{-j})} = \frac{\pi(\gamma_j = 1, \gamma_{-j})}{\pi(\gamma_j = 0, \gamma_{-j})} = \frac{\pi(|\gamma| = r_j + 1)/\binom{\bar{r}_n}{r_j+1}}{\pi(|\gamma| = r_j)/\binom{\bar{r}_n}{r_j}} \mathbf{1}_{(r_j+1 \leq \bar{r}_n)}$$

Notice

$$\frac{\pi(|\gamma| = r_j + 1)}{\pi(|\gamma| = r_j)} = \frac{\int_0^\infty \pi(X = r_j + 1|\lambda) \pi(\lambda) d\lambda}{\int_0^\infty \pi(X = r_j|\lambda) \pi(\lambda) d\lambda},$$

where $X \sim Pois(\lambda)$, $\lambda \sim Ga(a_n, b_n)$. The Poisson-Gamma mixture is a Negative Binomial distribution.

$$\frac{\pi(|\gamma| = r_j + 1)}{\pi(|\gamma| = r_j)} = \frac{\Gamma(r_j + 1 + a_n)/\Gamma(r_j + 2)}{\Gamma(r_j + a_n)/\Gamma(r_j + 1)} \frac{1}{1 + b_n} = \frac{r_j + a_n}{(r_j + 1)(b_n + 1)}$$

Also notice

$$\frac{\binom{\bar{r}_n}{r_j}}{\binom{\bar{r}_n}{r_j+1}} = \frac{\bar{r}_n! / (r_j! (\bar{r}_n - r_j)!)}{\bar{r}_n! / ((r_j + 1)! (\bar{r}_n - r_j - 1)!)} = \frac{r_j + 1}{\bar{r}_n - r_j}$$

Then, the odds is calculated:

$$\frac{p_j}{1-p_j} = \frac{r_j + a_n}{(r_j + 1)(b_n + 1)} \frac{r_j + 1}{\bar{r}_n - r_j} \mathbf{1}_{(r_j+1 \leq \bar{r}_n)} = \frac{(r_j + a_n) \mathbf{1}_{(r_j+1 \leq \bar{r}_n)}}{(b_n + 1)(\bar{r}_n - r_j)}.$$

So

$$p_j = \frac{(r_j + a_n) \mathbf{1}_{(r_j+1 \leq \bar{r}_n)}}{(b_n + 1) \bar{r}_n - b_n r_j + a_n}$$

The relevant terms from the likelihood is

$$\begin{aligned} L(h, \varsigma) &\propto e^{\sigma_x \sum_{i=1}^n h_i x_i - \sigma_y \sum_{i=1}^n h_i y_i - \frac{1}{2} \sigma_h \sum_{i=1}^n h_i^2} \\ &\propto e^{\theta_j (\sigma_x \sum_{i=1}^n z_{ij} x_i - \sigma_y \sum_{i=1}^n z_{ij} y_i - \sigma_h \sum_{i=1}^n z_{ij} h_{ij}) - \frac{1}{2} \theta_j^2 \sigma_h \sum_{i=1}^n z_{ij}^2} \end{aligned}$$

where $\sigma_x = \sigma_{11}^{-2} + \sigma_{2|1}^{-2} a_{21} (a_{21} + \beta)$, $\sigma_y = \sigma_{2|1}^{-2} a_{21}$, $\sigma_h = \sigma_{11}^{-2} + \sigma_{2|1}^{-2} a_{21}^2$, and $h_{ij} = h_i - z_{ij} \theta_j = \sum_{\ell \neq j} z_{i\ell} \theta_\ell$.

The conditional prior density is

$$\pi(\theta_j | \theta_{-j}) = p_j N(\theta_j | 0, \tau^2) + (1 - p_j) 1_{(\theta_j=0)}$$

So

$$\begin{aligned} \pi(\theta_j | \text{others}) &\propto L(h, \varsigma) \pi(\theta_j | \theta_{-j}) \\ &\propto e^{-\frac{1}{2} (\theta_j^2 \sigma_h \sum_{i=1}^n z_{ij}^2 - 2g_j \theta_j)} (p_j N(\theta_j | 0, \tau^2) + (1 - p_j) 1_{(\theta_j=0)}), \end{aligned}$$

where $g_j = \sigma_x \sum_{i=1}^n z_{ij} x_i - \sigma_y \sum_{i=1}^n z_{ij} y_i - \sigma_h \sum_{i=1}^n z_{ij} h_{ij}$.

Completing the squares yields

$$\pi(\theta_j | \text{others}) \propto p_j \frac{N(0 | 0, \tau^2)}{N(0 | E_j, V_j)} N(\theta_j | E_j, V_j) + (1 - p_j) 1_{(\theta_j=0)}$$

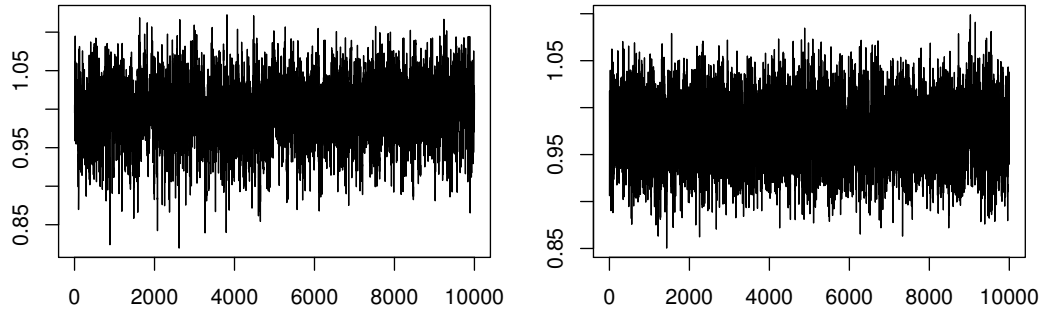
where $V_j^{-1} = \sigma_h \sum_{i=1}^n z_{ij}^2 + \tau^{-2}$ and $E_j = V_j g_j$.

Normalize the weights for sampling's purpose:

$$\theta_j | \text{others} \sim p_j^* N(\theta_j | E_j, V_j) + (1 - p_j^*) 1_{(\theta_j=0)}$$

A.5 Mixing of β and θ

For the benchmark case, starting value is set to be the true value. The mixing in Figure A.1 seems convergent for both cases. The corresponding plot is Figure 4.2.



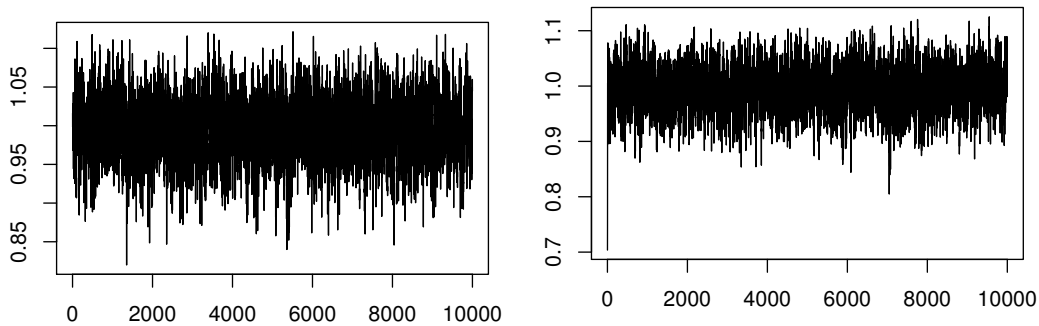
(a) Independent IV

(b) Dependent IV

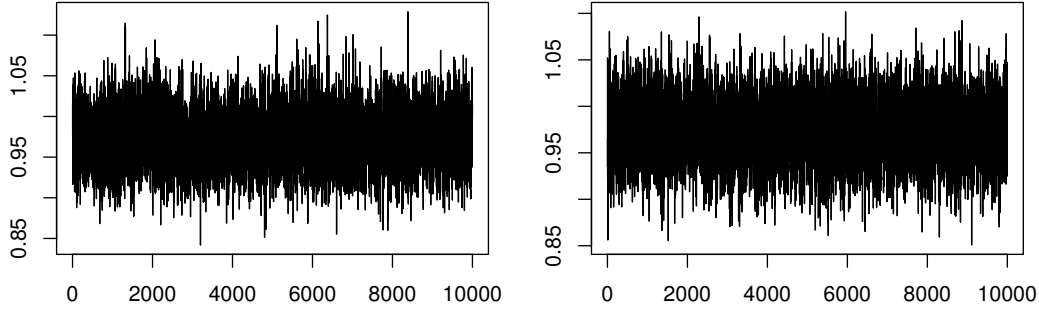
FIGURE A.1: Mixing of β , Starting Value=1

Two different starting values are examined. One is -1 and the other is 3. -1 and 3 are both away from 1. If we choose values more deviate from 1, the Gibbs sampler may encounter numerical difficulties as the probability evaluated at somewhere of a distant tail of a Normal distribution is almost 0, which produces numerical errors.

From Figure A.2, we can see that the effect of different starting values matters little. The Gibbs sampler quickly finds the right region with 50 burn-in draws.



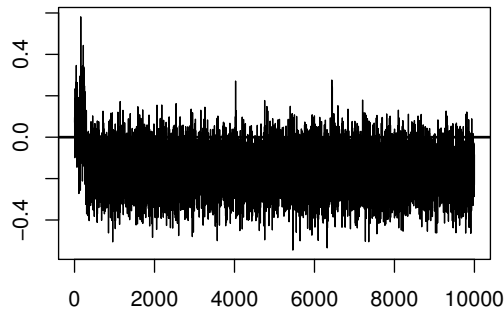
(a) Starting Value=3, in Independent IV (b) Starting Value=-1, in Independent IV



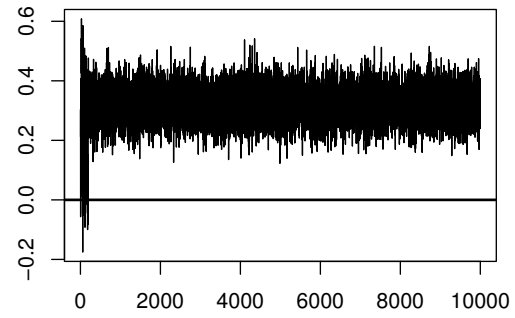
(c) Starting Value=3, in Dependent IV (d) Starting Value=-1, in Dependent IV

FIGURE A.2: Mixing of β at different starting values

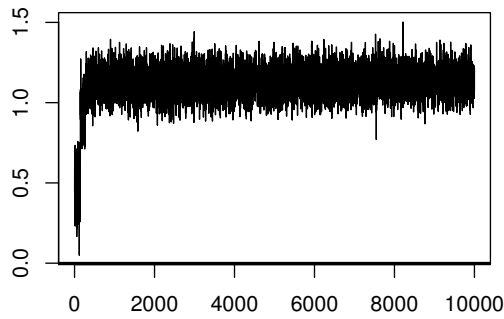
To examine the mixing of θ , we can look at the mixing of h . As the dimension of h depends on the sample size. We can randomly pick some observations and see the mixing of h_i . The mixing of h_i in both cases looks good too even though the beginning of chain is quite different from the later parts of the chain. This indicates that the Gibbs sampler has difficulty in finding the right region at beginning. But with more burn-in draws, the saved draws should be convergent throughout the chain.



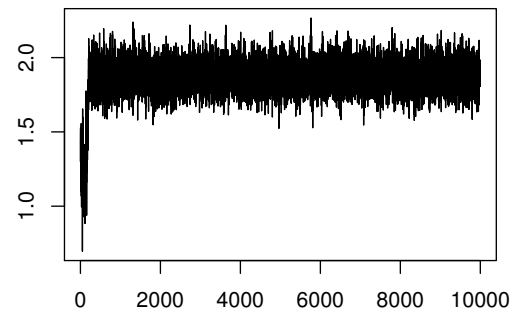
(a) h_{100} in Independent IV



(b) h_{100} in Dependent IV



(c) h_{200} in Independent IV



(d) h_{200} in Dependent IV

FIGURE A.3: Mixing of h_i

Bibliography

- Anderson, T., Kunitomo, N., and Matsushita, Y. “On the asymptotic optimality of the LIML estimator with possibly many instruments.” *Journal of Econometrics*, 157(2):191–204 (2010).
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. “Identification of causal effects using instrumental variables.” *Journal of the American statistical Association*, 91(434):444–455 (1996).
- Bai, J. and Ng, S. “Instrumental variable estimation in a data rich environment.” *Econometric Theory*, 26(06):1577–1606 (2010).
- Bekker, P. A. “Alternative approximations to the distributions of instrumental variable estimators.” *Econometrica: Journal of the Econometric Society*, 657–681 (1994).
- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. “Sparse models and methods for optimal instruments with an application to eminent domain.” *Econometrica*, 80(6):2369–2429 (2012).
- Caner, M. “Lasso-type GMM estimator.” *Econometric Theory*, 25(01):270–290 (2009).
- Chamberlain, G. and Imbens, G. “Hierarchical Bayes models with many instrumental variables.” (1996).
- Chan, J. C. and Tobias, J. L. “Priors and Posterior Computation in Linear Endogenous Variable Models with Imperfect Instruments.” *Journal of Applied Econometrics* (2014).
- Chao, J. C. and Phillips, P. C. “Posterior distributions in limited information analysis of the simultaneous equations model using the Jeffreys prior.” *Journal of Econometrics*, 87(1):49–86 (1998).
- Chao, J. C. and Swanson, N. R. “Consistent estimation with a large number of weak instruments.” *Econometrica*, 73(5):1673–1692 (2005).

- Conley, T. G., Hansen, C. B., McCulloch, R. E., and Rossi, P. E. “A semi-parametric Bayesian approach to the instrumental variable problem.” *Journal of Econometrics*, 144(1):276–305 (2008).
- Conley, T. G., Hansen, C. B., and Rossi, P. E. “Plausibly exogenous.” *Review of Economics and Statistics*, 94(1):260–272 (2012).
- Drèze, J. H. “Bayesian regression analysis using poly-t densities.” *Journal of Econometrics*, 6(3):329–354 (1977).
- Dreze, J. H. and Richard, J.-F. “Bayesian analysis of simultaneous equation systems.” *Handbook of econometrics*, 1:517–598 (1983).
- George, E. I. and McCulloch, R. E. “Variable selection via Gibbs sampling.” *Journal of the American Statistical Association*, 88(423):881–889 (1993).
- . “Approaches for Bayesian variable selection.” *Statistica sinica*, 7(2):339–373 (1997).
- Geyer, C. J. “Markov chain Monte Carlo maximum likelihood.” (1991).
- Geyer, C. J. and Thompson, E. A. “Annealing Markov chain Monte Carlo with applications to ancestral inference.” *Journal of the American Statistical Association*, 90(431):909–920 (1995).
- Hansen, C., Hausman, J., and Newey, W. “Estimation with many instrumental variables.” *Journal of Business & Economic Statistics*, 26(4) (2008).
- Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., and Swanson, N. R. “Instrumental variable estimation with heteroskedasticity and many instruments.” *Quantitative Economics*, 3(2):211–255 (2012).
- Hoogerheide, L., Kleibergen, F., and van Dijk, H. K. “Natural conjugate priors for the instrumental variables regression model applied to the Angrist–Krueger data.” *Journal of Econometrics*, 138(1):63–103 (2007).
- Jiang, W. “Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities.” *The Annals of Statistics*, 35(4):1487–1511 (2007).
- Kapetanios, G. and Marcellino, M. “Factor-GMM estimation with large sets of possibly weak instruments.” *Computational Statistics & Data Analysis*, 54(11):2655–2675 (2010).
- Kleibergen, F. and Zivot, E. “Bayesian and classical approaches to instrumental variable regression.” *Journal of Econometrics*, 114(1):29–72 (2003).

- Koop, G., Leon-Gonzalez, R., and Strachan, R. “Bayesian model averaging in the instrumental variable regression model.” *Journal of Econometrics*, 171(2):237–250 (2012).
- Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. “Penalized regression, standard errors, and Bayesian lassos.” *Bayesian Analysis*, 5(2):369–411 (2010).
- Lancaster, T. *An introduction to modern Bayesian econometrics*. Blackwell Oxford (2004).
- Laurent, B. and Massart, P. “Adaptive estimation of a quadratic functional by model selection.” *Annals of Statistics*, 1302–1338 (2000).
- Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. “A significance test for the lasso.” *Annals of statistics*, 42(2):413 (2014).
- Lopes, H. F. and Polson, N. G. “Bayesian instrumental variables: priors and likelihoods.” *Econometric Reviews*, 33(1-4):100–121 (2014).
- Pollard, D. “Empirical processes: theory and applications.” In *NSF-CBMS regional conference series in probability and statistics*, i–86. JSTOR (1990).
- Ročková, V. and George, E. I. “Emvs: The em approach to bayesian variable selection.” *Journal of the American Statistical Association*, 109(506):828–846 (2014).
- Rossi, P. E., Allenby, G. M., and McCulloch, R. E. *Bayesian statistics and marketing*. J. Wiley & Sons (2005).
- Stock, J. H., Wright, J. H., and Yogo, M. “A survey of weak instruments and weak identification in generalized method of moments.” *Journal of Business & Economic Statistics*, 20(4) (2002).
- Stock, J. H. and Yogo, M. “Testing for weak instruments in linear IV regression.” *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 1 (2005).
- Varian, H. R. “Big data: New tricks for econometrics.” *The Journal of Economic Perspectives*, 3–27 (2014).