Fully Bayesian endogenous variable estimation with many instrumental variables

by

Ethan Schubert

B.S., Wheaton College (IL), 2022

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Co-Major Professor
Gyuhyeong Goh

Approved by:

Co-Major Professor
Christopher Vahl

Copyright

Abstract

Endogeneity remains a thorny problem in applications of regression analysis, where the correlation with random error may introduce large bias into the parameter estimates. Instrumental variables are an elegant solution to the problem, but only when the instrumental variables affect the response solely through the endogenous variable. If any subset of the instrumental variables are included directly in the response model, bias is reintroduced through these invalid instruments. In practice, it is impossible to know for certain which instruments may be invalid, but much work has been done to estimate the validity of candidate instruments through penalized regression methods. We introduce a Bayesian alternative to these methods with oracle properties which also accounts for model uncertainty through Bayesian model averaging. Our model contains a sparsity assumption on the invalid instruments, specifically that the invalid instruments are less than half of the total set of candidate instruments, and we introduce an alternative construction of the Gibbs sampler to account for the sparsity constraint. Our estimator demonstrates MSE comparable to oracle estimators in several simulation studies. We also apply the estimator to a real-world dataset on global trade to identify invalid instruments.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

A common and useful technique for estimating the causal effect of an endogenous variable on the outcome of interest is the use of instrumental variables. Instrumental variables (IVs) are covariates which are not of primary interest to the researcher but are useful for estimating the endogenous effect. Specifically, ordinary least squares regression is biased under endogeneity, and IVs can be used to produce unbiased estimates of the endogenous effect. This method is known as two-stage least squares regression (TSLS). Essentially, endogeneity is removed by first regressing the endogenous variable on a set of IVs, and then by regressing the reponse on the resulting estimate instead of the original endogenous variable. To illustrate, we write the classical form of the IV problem as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ x_i &= \mathbf{z}_i^T \boldsymbol{\alpha} + \eta_i, \quad i = 1, \ldots, n \end{aligned} \tag{1.1}$$

where y_i is the i^{th} observation of the response variable, x_i is the i^{th} observation of the endogenous variable of interest, and $\mathbf{z}_i^T$ are the i^{th} observations of p instrumental variables for x. Since x is endogenous, by definition $Cov(x_i, \epsilon_i) \neq 0$ for some i. However, it is assumed that $Cov(\mathbf{z}_i, \epsilon_i) = 0$ and $Cov(\mathbf{z}_i, \epsilon_i) = 0$ for all i.

The TSLS estimator can then be constructed as follows. Let $\mathbf{Z}$ be the matrix of ob-

servations of the instrumental variables and $\mathbf{P_Z}$ be the projection matrix of $\mathbf{Z}$. Then, to construct the TSLS estimator for β_1, we first construct $\hat{\mathbf{x}} = \mathbf{P_Z x}$. Note that $Cov(\hat{\mathbf{x}}, \boldsymbol{\epsilon}) = 0$ since $Cov(\mathbf{z}_i, \epsilon_i) = 0$ for all i. Now, we regress $\mathbf{y}$ on $\hat{\mathbf{x}}$ to obtain the TSLS estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$: $\hat{\boldsymbol{\beta}}_{tsls} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ where $\mathbf{X} = [\mathbf{1}, \hat{\mathbf{x}}]$. The TSLS estimator is a consistent estimator for $\boldsymbol{\beta}$, solving the IV problem as given in (1.1). This method, however, assumes that the effect of $\mathbf{Z}$ on $\mathbf{y}$ is entirely expressed through $\hat{\mathbf{x}}$. In other words, no element of $\mathbf{Z}$ can appear in the model for $\mathbf{y}$. If this does not hold, the TSLS estimator just described is no longer a consistent estimator of $\boldsymbol{\beta}$. We modify the previous model for this situation as follows. Let $\boldsymbol{\gamma}$ be a $p \times 1$ vector corresponding to the effect of the instrumental variables on y_i. Then the new model becomes

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \mathbf{z}_i^T\boldsymbol{\gamma} + \epsilon_i \\ x_i &= \mathbf{z}_i^T\boldsymbol{\alpha} + \eta_i, \quad i = 1, \ldots, n \end{aligned} \tag{1.2}$$

When $\gamma_j = 0$, the corresponding instrumental variable $\mathbf{z}_j$ is called a valid instrumental variable. This instrumental variable only affects $\mathbf{y}$ through $\mathbf{x}$, satisfying the assumptions of the TSLS. Contrastingly, when $\gamma_j \neq 0$, the corresponding instrumental variable is an invalid instrumental variable. If $\boldsymbol{\gamma}$ is known, we can modify the TSLS estimator to account for the effect of the invalid IVs using the generalized method of moments. For this method, we first center the data and compute $\hat{\mathbf{x}} = \mathbf{P_Z x}$ as before. Then, let $\mathbf{Z}_2$ be a submatrix of $\mathbf{Z}$ corresponding to the invalid IVs, and define $\mathbf{M}_{\mathbf{Z}_2} = \mathbf{I}_n - \mathbf{P}_{\mathbf{Z}_2}$, where $\mathbf{P}_{\mathbf{Z}_2}$ is the projection matrix of $\mathbf{Z}_2$. Then, the generalized method of moments estimator is $\hat{\beta}_{_{oracle}} = \left(\hat{\mathbf{x}}^T\mathbf{M}_{\mathbf{Z}_2}\hat{\mathbf{x}}\right)^{-1}\left(\hat{\mathbf{x}}^T\mathbf{M}_{\mathbf{Z}_2}\mathbf{y}\right)$ which accounts for the effects of the invalid IVs and is a consistent estimator of $\boldsymbol{\beta}$.

Unfortunately, $\boldsymbol{\gamma}$ is not known in practice, and thus it is impossible to specify $\mathbf{Z}_2$ to construct $\hat{\beta}_{_{oracle}}$. We will refer to this as the unknown invalid instrumental variables (UIIV) problem to distinguish it from the invalid IV problem with known $\boldsymbol{\gamma}$. Kang et al (2016)[1] establish an important condition for the estimation of UIIVs. They show that unique esti-

mation of β_1 is possible when the number of invalid instruments is less than 50% of the set of instruments used. Under this requirement, estimators have been developed for the UIIV problem with oracle properties, such as sisVIVE in Kang et. al (2016)[1] and the median estimators in Windmeijer et. al. (2019)[2]. These estimation methods are both penalized regression methods and are effective under a properly chosen tuning parameter. However, there is no method to choose an optimal tuning parameter, and thus in practice the effectiveness of these estimators cannot be guaranteed.

Goh and Yu (2022)[3] attempt to address the problem of model uncertainty with a quasi-Bayesian estimator. However, although their estimator also exhibits oracle properties, the method does not entirely account for all model uncertainty. This paper introduces a fully Bayesian method for endogenous effect estimation in the UIIV setting which accounts for all model uncertainty. Our proposed method also demonstrates oracle properties in simulation studies. Following the results in Kang et al (2016)[1] and Windmeijer et al (2019)[2], we require that the number of invalid instrumental variables be less than 50% of the total number of instrumental variables for our method. We formally define the model and priors used in the next section.

Chapter 2

Bayesian model set up

As before, we set up the UIIV problem as follows. Let y_i be the i^{th} observation of the outcome variable, x_i be the i^{th} observation of the endogenous variable, and $\mathbf{z}_i^T$ be the i^{th} observations of p instrumental variables. Then the regression model becomes

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \mathbf{z}_i^T \boldsymbol{\gamma} + \epsilon_i \\ x_i &= \mathbf{z}_i^T \boldsymbol{\alpha} + \eta_i \end{aligned} \tag{2.1}$$

In this framework, β_0, β_1, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ are unknown parameters where β_0 is the intercept, β_1 is the effect of the endogenous variable, and $\boldsymbol{\alpha}$ is the effect of the instrumental variables on x_i. $\boldsymbol{\gamma}$ is a $p \times 1$ vector corresponding to the effect of the invalid instruments on y_i. An instrumental variable $\mathbf{z}_j$ is invalid if $\gamma_j \neq 0$. If $\gamma_j = 0$, then $\mathbf{z}_j$ is a valid IV. We require $\boldsymbol{\gamma}$ to be sparse, where more than $p/2$ elements of $\boldsymbol{\gamma}$ are zero. This is equivalent to requiring fewer than half of the instrumental variables to be invalid, as in Kang et al (2016)[1] and Windmeijer et al (2019)[2]. Additionally, ϵ_i and η_i denote random errors, where $(\epsilon_i, \eta_i)^T$ and $(\epsilon_j, \eta_j)^T$ are independent for all $i \neq j$. We further assume that $(\epsilon_i, \eta_i)^T$ has the following structure:

$$\begin{bmatrix} \epsilon_i \\ \eta_i \end{bmatrix} \sim N_2(0, \boldsymbol{\Sigma}), \text{ where } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

with $\sigma_{12} = \sigma_{21} \neq 0$. Lastly, we assume that $\mathbf{z}_i^T$ and $(\epsilon_i, \eta_i)^T$ are independent.

To create a Bayesian specification of the UIIV problem, we assign prior distributions to all unobserved quantities. For the prior distribution of $\boldsymbol{\gamma}$, we need to assign a distribution which retains the sparsity constraint on $\boldsymbol{\gamma}$. The sparsity of $\boldsymbol{\gamma}$ is important, because identifying non-zero γ_j is equivalent to selecting invalid IVs. Since we require invalid IVs to consist of fewer than 50% of the set of IVs to estimate the endogenous effect, our model would be inappropriate if $\boldsymbol{\gamma}$ was not sparse. To ensure sparsity, we will introduce a set of independent and identically distributed Bernoulli random variables s_j where $j \in 1, \ldots, p$. We then assign the prior distribution of γ_j using an exact-zero spike-slab prior in the following hierarchical model:

$$\begin{aligned} \gamma_j | s_j &\sim (1 - s_j)\delta_0 + s_j \phi(0, \tau_\gamma^2) \\ s_j &\overset{iid}{\sim} Ber(\omega) \end{aligned} \tag{2.2}$$

where $\omega \in (0, 1)$ and δ_0 is the unit impulse function at 0. Additionally, $\phi(0, \tau_\gamma^2)$ denotes the density of a Normal distribution with mean 0 and variance τ_γ^2. We choose the hyperparameter τ_γ^2 to be large such that the prior distribution of γ_j is non-informative.

For the remaining parameters, we assign non-informative conjugate priors to ease computation:

$$\begin{aligned} \boldsymbol{\alpha} &\sim N_p(0, \tau_\alpha^2 \mathbf{I}_p) \\ (\beta_0, \beta_1)^T &\sim N_2(0, \tau_\beta^2 \mathbf{I}_2) \\ \boldsymbol{\Sigma} &\sim invWishart\,(\nu, \boldsymbol{\Psi}) \end{aligned} \tag{2.3}$$

As before, the parameters τ_α^2 and τ_β^2 are chosen to be large so that the prior distributions are non-informative.

Our objective is to sample from $p(\beta_1 | \mathbf{x}, \mathbf{y}, \mathbf{Z})$, the marginal distribution of β_1 given the data. The standard Bayesian approach to accomplish this is through the construction of a

Gibbs sampler. After a sufficient number of iterations, the Gibbs sampler is able to sample from $p(\beta_1|\mathbf{x}, \mathbf{y}, \mathbf{Z})$, starting from the full conditional distributions of the parameters and initial values of the parameters. In this case, the full conditional distributions required are:

$$\begin{aligned}
\boldsymbol{\alpha} &\sim p(\boldsymbol{\alpha}|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \\
\beta_0 &\sim p(\beta_0|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \beta_1, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \\
\beta_1 &\sim p(\beta_1|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \\
\boldsymbol{s} &\sim p(\boldsymbol{s}|\boldsymbol{\gamma}) \\
\boldsymbol{\gamma} &\sim p(\boldsymbol{\gamma}|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{s}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &\sim p(\boldsymbol{\Sigma}|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})
\end{aligned} \tag{2.4}$$

To derive the full conditionals, we can use Bayes' theorem. For a parameter θ with hyperparameter τ, such that $\theta \sim p(\theta|\tau)$, and data $\mathbf{D}$, we can express Bayes' theorem as follows:

$$p(\theta|\mathbf{D}) = \frac{f(\mathbf{D}|\theta, \tau)p(\theta|\tau)}{\int f(\mathbf{D}|\theta, \tau)p(\theta|\tau)d\theta} \tag{2.5}$$

where $f(\mathbf{D}|\theta, \tau)$ is the likelihood function of the data. Furthermore, since the denominator is free of θ, when deriving the posterior distribution we can simplify Bayes' theorem to a proportional form

$$p(\theta|\mathbf{D}) \propto f(\mathbf{D}|\theta, \tau)p(\theta|\tau) \tag{2.6}$$

and use the fact that the distribution must integrate to 1 to derive the exact form of the density function. Once the forms of these conditional distributions are known, we can initiate the Gibbs sampler with starting values for $\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{s}$ and $\boldsymbol{\Sigma}$ (denoted as $\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\gamma}}_0, \tilde{\mathbf{s}}_0$ and $\tilde{\boldsymbol{\Sigma}}_0$) to generate a value $\tilde{\boldsymbol{\alpha}}_1$ from the full conditional distribution of $\boldsymbol{\alpha}$. Then, the sampled value $\tilde{\boldsymbol{\alpha}}_1$ is used to generate $\tilde{\beta}_{01}$ from the full conditional distribution of β_0. We then continue to sucessively sample from the full conditionals, using the most recently sampled value for given parameters. Thus the sampled parameter values for the t^{th} iteration are $\tilde{\boldsymbol{\alpha}}_t, \tilde{\beta}_{0t}, \tilde{\beta}_{1t}, \tilde{\boldsymbol{\gamma}}_t, \tilde{\mathbf{s}}_t$ and $\tilde{\boldsymbol{\Sigma}}_t$. The strength of the Gibbs sampler is that after a sufficiently large number of iterations

N, the distribution of the sampled values of β_1 converge to the marginal distribution of β_1 given the data, averaged over the other parameters.

Although at this point the approach should appear straightforward, unfortunately the model construction introduces a difficulty in sampling $\boldsymbol{\gamma}$ and $\mathbf{s}$ which prevents us from using the Gibbs sampler described previously. In the next section, we further describe this problem and propose a modified Gibbs sampler.

Chapter 3

Posterior inference

Our construction of the model preserves the sparsity of $\boldsymbol{\gamma}$ which is crucial to the estimation of the endogenous effect. However, this model construction introduces a difficulty into the Gibbs sampler through the sampling distribution of $\mathbf{s}$. To see why, we will examine the simpler case of the posterior distribution of s_j. Using Bayes' theorem, the posterior distribution of s_j is proportional to the product of the prior distributions of $\gamma_j|s_j$ and s_j. That is, $p(s_j|\gamma_j) \propto p(\gamma_j|s_j)p(s_j)$. Expressing the prior distribution in their explicit forms, we have

$$p(s_j|\gamma_j) \propto \left[(1-s_j)I(\gamma_j = 0) + s_j\phi(\gamma_j|0,\tau_\gamma^2)\right]\omega^{s_j}(1-\omega)^{1-s_j} \tag{3.1}$$

Then, the proportional form of the success probability is given by $p(s_j = 1|\gamma_j) \propto \phi(\gamma_j|0,\tau^2)\omega$ and the failure probability by $p(s_j = 0|\gamma_j) \propto I(\gamma_j = 0)(1-\omega)$. This implies that

$$s_j|\gamma_j \sim Ber\left(\frac{\phi(\gamma_j|0,\tau^2)\omega}{\phi(\gamma_j|0,\tau^2)\omega + I(\gamma_j = 0)(1-\omega)}\right) \tag{3.2}$$

Now, suppose that the true value of γ_j is 0, but for some iteration t of the Gibbs sampler, the sampled value $\tilde{\gamma}_{jt}$ is close to (but not equal to) 0. Then the conditional distribution simplifies to $s_j|\tilde{\gamma}_{jt} \sim Ber(\phi(\gamma_j|0,\tau^2)\omega[\phi(\gamma_j|0,\tau^2)\omega]^{-1}) = Ber(1) = 1$. This restricts the parameter space of s_j from $\{0,1\}$ to $\{1\}$, reducing the dimension of the parameter space

for all future iterations of the Gibbs sampler. Furthermore, if this occurs, γ_j will never be estimated to be 0, violating the sparsity condition of the model. This dimension-changing problem prevents the Gibbs sampler from converging to the correct stationary distribution and thus the traditional approach is not valid (Green 1995)[4].

To resolve this problem, we propose a modification to the Gibbs sampler by sampling (γ_j, s_j) jointly. That is, we will sample $s_j|\mathbf{s}_{-j}, \boldsymbol{\gamma}_{-j}$ and $\gamma_j|\mathbf{s}, \boldsymbol{\gamma}_{-j}$ where $\mathbf{s}_{-j}, \boldsymbol{\gamma}_{-j}$ are the vectors $\mathbf{s}$ and $\boldsymbol{\gamma}$ without their j^{th} elements. This modification preserves the dimension of $\mathbf{s}$ and $\boldsymbol{\gamma}$ throughout the Gibbs sampler, which allows us to sample from the full conditionals. The details of this derivation are given later. Now, we give the formal definition of our Gibbs sampler:

1. First, we set initial values for our parameters, denoted by $\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\gamma}}_0, \tilde{\mathbf{s}}_0$, and $\tilde{\boldsymbol{\Sigma}}_0$, respectively. We then sample from the following conditional distributions:
2. $\boldsymbol{\alpha} \sim p(\boldsymbol{\alpha}|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$.
3. $\beta_0 \sim p(\beta_0|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \beta_1, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$.
4. $\beta_1 \sim p(\beta_1|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$.
5. Then for $j \in \{1, \ldots, p\}$ we sample from

   $s_j \sim p(s_k|\boldsymbol{\gamma}_{-j}, \mathbf{s}_{-j})$ and

   $\gamma_j \sim p(\gamma_j|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{-j}, \mathbf{s}, \boldsymbol{\Sigma})$.
6. Then we sample $\boldsymbol{\Sigma} \sim p(\boldsymbol{\Sigma}|\mathbf{x}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
7. Repeat steps 2-6 for a sufficiently large number of iterations.

In our simulations, we have found the sampler converges relatively quickly, with 3000 iterations being a sufficient burn-in period. For initial values, we recommend using a standard OLS estimate for $\tilde{\boldsymbol{\beta}}_0$. Additionally, we set $\tilde{\boldsymbol{\Sigma}}_0 = \mathbf{I}_2$, $\tilde{\boldsymbol{\gamma}}_0 = 0$ and $\tilde{\mathbf{s}}_0 = 0$.

Chapter 4

Derivation of full conditionals

Now we will specify the forms of the conditional distributions discussed previously. First, we will establish the forms of the likelihood functions of $\mathbf{x}$ and $\mathbf{y}$. To simplify our notation, we will rewrite the model in matrix form as

$$\begin{aligned} \mathbf{y} &= \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ \mathbf{x} &= \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\eta} \end{aligned} \tag{4.1}$$

where $\mathbf{y}$ and $\mathbf{x}$ are $n \times 1$ vectors, $\mathbf{Z}$ is an $n \times p$ matrix, $\mathbf{D} = [\mathbf{1}_n, \mathbf{x}, \mathbf{Z}]$, and $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^T)^T$. Now, recall that $(\epsilon_i, \eta_i)^T \sim N_2(0, \boldsymbol{\Sigma})$. Thus, the conditional distributions $\epsilon_i|\eta_i$ and $\eta_i|\epsilon_i$ are $\epsilon_i|\eta_i \sim N(0, \sigma_{11.2})$ and $\eta_i|\epsilon_i \sim N(0, \sigma_{22.1})$ where $\sigma_{11.2} = \sigma_{11} - \sigma_{12}\sigma_{22}^{-1}\sigma_{21}$ and $\sigma_{22.1} = \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}$. Also, recall that ϵ_i and ϵ_j are independent for $i \neq j$. Likewise, η_i and η_j are independent for $i \neq j$. Then we can write $\boldsymbol{\epsilon}|\boldsymbol{\eta} \sim N(0, \sigma_{11.2}\mathbf{I}_n)$ and $\boldsymbol{\eta}|\boldsymbol{\epsilon} \sim N(0, \sigma_{22.1}\mathbf{I}_n)$. This leads to the following form of the likelihood functions for $\mathbf{x}$ and $\mathbf{y}$:

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\epsilon}) &= \phi\left(\mathbf{Z}\boldsymbol{\alpha} + \sigma_{21}\sigma_{11}^{-1}\boldsymbol{\epsilon}, \sigma_{22.1}\mathbf{I}_n\right) \\ f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) &= \phi\left(\mathbf{D}\boldsymbol{\theta} + \sigma_{12}\sigma_{22}^{-1}\boldsymbol{\eta}, \sigma_{11.2}\mathbf{I}_n\right), \end{aligned} \tag{4.2}$$

where ϕ is the Normal density function with the specified mean and covariance matrix parameters. We will use the likelihood functions of $\mathbf{x}$ and $\mathbf{y}$ in deriving the full conditional

distributions of $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, and $\boldsymbol{\Sigma}$.

4.1 The Posterior Distribution of $\boldsymbol{\alpha}$

The posterior distribution of $\boldsymbol{\alpha}$ is proportional to the likelihood function of $\mathbf{x}$ times the prior distribution of $\boldsymbol{\alpha}$, by Bayes theorem:

$$p(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{D}, \boldsymbol{\theta}) \propto f(\mathbf{x}|\boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\epsilon}) p(\boldsymbol{\alpha}) \tag{4.3}$$

We will show that the posterior distribution of $\boldsymbol{\alpha}$ is Normal by expressing it in terms of the kernel of a Normal density function. Thus, in our derivation, we will ignore the constant terms of the distribution functions and express them solely in terms containing $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{D}, \boldsymbol{\theta}) \propto exp\left(-\frac{1}{2\sigma_{22.1}}||\mathbf{x} - \mathbf{Z}\boldsymbol{\alpha} - \sigma_{21}\sigma_{11}^{-1}\boldsymbol{\epsilon}||^2\right) exp\left(-\frac{1}{2\tau_\alpha^2}||\boldsymbol{\alpha}||^2\right).$$

We then combine the exponential terms and rewrite $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{D}\boldsymbol{\theta}$ so that the exponential term becomes $exp(-\frac{1}{2\sigma_{22.1}}(||-\mathbf{Z}\boldsymbol{\alpha} + \mathbf{x} - \sigma_{21}\sigma_{11}^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\theta})||^2 + \frac{\sigma_{22.1}}{\tau_\alpha^2}||\boldsymbol{\alpha}||^2)$. Next, we expand the first inner product inside the exponential:

$$||-\mathbf{Z}\boldsymbol{\alpha} + \mathbf{x} - \sigma_{21}\sigma_{11}^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\theta})||^2 = (\mathbf{Z}\boldsymbol{\alpha})^T\mathbf{Z}\boldsymbol{\alpha} - 2(\mathbf{Z}\boldsymbol{\alpha})^T(\mathbf{x} - \sigma_{21}\sigma_{11}^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\theta}) + c$$

where $c = ||(\mathbf{x} - \sigma_{21}\sigma_{11}^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\theta})||^2$. Notice that c does not depend on $\boldsymbol{\alpha}$. Thus, we can drop it from the proportional form of the full conditional distribution. Furthermore, since $(\mathbf{Z}\boldsymbol{\alpha})^T(\mathbf{Z}\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\alpha}$, we combine this term with the other quadratic form $\boldsymbol{\alpha}^T\boldsymbol{\alpha}$ to rewrite the exponential term as $(-\frac{1}{2\sigma_{22.1}}(\boldsymbol{\alpha}^T(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_{22.1}}{\tau_\alpha^2}\mathbf{I}_p)\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\mathbf{Z}^T(\mathbf{x} - \sigma_{21}\sigma_{11}^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\theta}))$.

Now, we will rewrite the expression in the familiar form of the kernel of a multivariate Normal density function. To do this, first note that for a $p \times p$ matrix $\mathbf{A}$ and a $p \times 1$ vector

$\mathbf{b}$, $\boldsymbol{\alpha}^T\mathbf{A}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\mathbf{b} = (\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})^T\mathbf{A}(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}) + \mathbf{c}$, where $\bar{\boldsymbol{\alpha}} = \mathbf{A}^{-1}\mathbf{b}$ and $\mathbf{c}$ is a constant. We set $\mathbf{A} = (\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_{22.1}}{\tau_\alpha^2}\mathbf{I}_p)$ and $\mathbf{b} = \mathbf{Z}^T(\mathbf{x} - \sigma_{21}\sigma_{11}^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\theta})$ and rewrite the exponential term:

$$\frac{1}{2\sigma_{22.1}}\left(\boldsymbol{\alpha}^T\mathbf{A}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\mathbf{b}\right) = \frac{1}{2\sigma_{22.1}}\left(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\right)^T\mathbf{A}(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}) + \frac{1}{2\sigma_{22.1}}\mathbf{c},$$

where $\bar{\boldsymbol{\alpha}} = (\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_{22.1}}{\tau_\alpha^2}\mathbf{I}_p)^{-1}\mathbf{Z}^T(\mathbf{x} - \sigma_{21}\sigma_{11}^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\theta}))$. Since the last term does not depend on $\boldsymbol{\alpha}$, we can remove it from the expression. Hence, we can rewrite the posterior distribution in the form of a Normal distribution:

$$p(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{D}, \boldsymbol{\theta}) \propto exp\left(-\frac{1}{2\sigma_{22.1}}(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})^T(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_{22.1}}{\tau_\alpha^2}\mathbf{I}_p)(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})\right) \tag{4.4}$$

Therefore, the posterior distribution of $\boldsymbol{\alpha}$ follows a Normal distribution with mean $\bar{\boldsymbol{\alpha}}$ and variance $\sigma_{22.1}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_{22.1}}{\tau_\alpha^2}\mathbf{I}_p\right)^{-1}$.

4.2 The Posterior Distribution of $\boldsymbol{\theta}$

To derive the distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we will first express them in a different form to ease notation. Recall that $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^T)^T$ and that

$$\begin{aligned}
\beta_0 &\sim N(0, \tau_\beta^2) \\
\beta_1 &\sim N(0, \tau_\beta^2) \\
\gamma_j|s_j &\sim (1 - s_j)\delta_0 + s_j\phi(0, \tau_\gamma^2) \text{ and} \\
s_j &\overset{iid}{\sim} Ber(\omega) \quad \text{for } j = 1, \ldots, p.
\end{aligned}$$

Now let $k \in \{0, \ldots, p+1\}$ so that $\theta_0 = \beta_0, \theta_1 = \beta_1$, and $\theta_k = \gamma_{j-1}$ for $k \geq 2$. Then, we define $\tau_{\theta_0}^2 = \tau_{\theta_1}^2 = \tau_\beta^2$ and $\tau_{\theta_k}^2 = \tau_\gamma^2$ for $k \geq 2$. Lastly, let $s_0 = s_1 = 1$ and $s_k \overset{iid}{\sim} Ber(\omega)$ for $k \in \{2, \ldots, p+1\}$. Then, we can write

$$\theta_k|s_k \sim (1 - s_k)\delta_0 + s_k\phi(\theta_k|0, \tau_{\theta_k}^2) \tag{4.5}$$

for notational convenience.

4.2.1 The Posterior Distribution of s_k

We will first derive the conditional distribution of s_k for $k \geq 2$, since s_0 and s_1 are set to 1. For $k \geq 2$, we will express the posterior distribution of s_k as the marginal density of the joint posterior of (s_k, θ_k).

$$p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) = \int p(s_k, \theta_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) d\theta_k \tag{4.6}$$

where $\boldsymbol{\theta}_{-k}$ and $\mathbf{s}_{-k}$ are the vectors $\boldsymbol{\theta}$ and $\mathbf{s}$ without their k^{th} elements. Using Bayes' theorem, we can express the posterior in the following form:

$$p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto \int f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) p(s_k, \theta_k) d\theta_k \tag{4.7}$$

Then, since the joint density is equal to the product of the marginal and conditional densities, we can rewrite the distribution as

$$p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto \int f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) p(\theta_k|s_k) p(s_k) d\theta_k \tag{4.8}$$

In this form, the explicit forms of the density functions on the right-hand side of the expression are known. We rewrite $p(\theta_k|s_k)$ in its explicit form and move $p(s_k)$ outside of the integral since it does not depend on θ_k:

$$p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto p(s_k) \int f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) [(1 - s_k)\delta_0 + s_k \phi(\theta_k|0, \tau^2_{\theta_k})] d\theta_k \tag{4.9}$$

Let $g(\theta_k)$ denote the integral term in the above expression. Notice that this integral can be

separated into two integral expressions, where $g(\theta_k) = g_1(\theta_k) + g_2(\theta_k)$. That is,

$$g(\theta_k) = (1 - s_k) \int f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) \delta_0 d\theta_k + s_k \int f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) \phi(\theta_k|0, \tau^2_{\theta_k}) d\theta_k.$$

We will evaluate the two integral expressions separately, starting with $g_1(\theta_k)$. Since δ_0 is a function of θ_k, the first integral evaluates to the likelihood when $\theta_k = 0$:

$$g_1(\theta_k) = (1 - s_k) \int f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) \delta_0 d\theta_k = (1 - s_k) f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}_{-k}, \boldsymbol{\eta}, \theta_k = 0). \tag{4.10}$$

Evaluating the second integral expression is more difficult. First, recall that $\phi(\theta_k|0, \tau^2_{\theta_k}) = (2\pi\tau^2_{\theta_k})^{-\frac{1}{2}} exp(-\theta^2_k/2\tau^2_{\theta_k})$ and that

$$f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) = (2\pi\sigma_{11.2})^{-\frac{n}{2}} exp(-||\mathbf{y} - \mathbf{D}\boldsymbol{\theta} - \sigma_{12}\sigma^{-1}_{22}\boldsymbol{\eta}||^2/2\sigma_{11.2}).$$

Since $p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto g(\theta_k) = g_1(\theta_k) + g_2(\theta_k)$, $g_2(\theta_k)$ is not proportional to the full conditional distribution of s_k. Thus, in this case we cannot drop the constant terms $(2\pi\tau^2_{\theta_k})^{-\frac{1}{2}}$ and $(2\pi\sigma_{11.2})^{-\frac{n}{2}}$ from the expression. However, since neither term depends on θ_k, we can move these terms outside of the integral. Let $c = (2\pi\tau^2_{\theta_k})^{-\frac{1}{2}}(2\pi\sigma_{11.2})^{-\frac{n}{2}}$. Additionally, we will write $\mathbf{D}\boldsymbol{\theta} = \mathbf{D}_{-k}\boldsymbol{\theta}_{-k} + \mathbf{d}_k\theta_k$, where $\mathbf{d}_k$ is the $k + 1^{th}$ column of $\mathbf{D}$ and $\mathbf{D}_{-k}$ is the matrix $\mathbf{D}$ without the $k + 1^{th}$ column. For example, note that $\mathbf{D}$ begins with a column of ones corresponding to $\theta_0 = \beta_0$. We denote this first column as $\mathbf{d}_0$. Also, for notational convenience let $\mathbf{y}^* = \mathbf{y} - \mathbf{D}_{-k}\boldsymbol{\theta}_{-k} - \sigma_{12}\sigma^{-1}_{22}\boldsymbol{\eta}$. We can then express $g_2(\theta_k)$ in a simplified form:

$$g_2(\theta_k) = s_k c \int exp\left[-\frac{1}{2\sigma_{11.2}}\left(||\mathbf{y}^* - \mathbf{d}_k\theta_k||^2 + \frac{\sigma_{11.2}}{\tau^2_{\theta_k}}\theta^2_k\right)\right] d\theta_k \tag{4.11}$$

Next, we expand the quadratic form $||\mathbf{y}^* - \mathbf{d}_k\theta_k||^2 = ||\mathbf{y}^*||^2 - 2\mathbf{y}^{*T}\mathbf{d}_k\theta_k + \theta^2_k||\mathbf{d}_k||^2$. Then, we move the constant term containing $||\mathbf{y}^*||^2$ outside the integral and define $c^* = c \cdot exp(-||\mathbf{y}^*||^2/2\sigma_{11.2})$. Then (4.11) becomes

$$g_2(\theta_k) = s_k c^* \int exp\left[-\frac{1}{2\sigma_{11.2}}\left(-2\mathbf{y}^{*T}\mathbf{d}_k\theta_k + \theta_k^2||\mathbf{d}_k||^2 + \frac{\sigma_{11.2}}{\tau_{\theta_k}^2}\theta_k^2\right)\right] d\theta_k \tag{4.12}$$

As in the derivation of the posterior distribution of $\boldsymbol{\alpha}$, we will express the exponential terms in centered form. Let $A = (||\mathbf{d}_k||^2 + \sigma_{11.2}/\tau_{\theta_k}^2)$ and $b = \mathbf{y}^{*T}\mathbf{d}_k$. Then, $A\theta_k^2 - 2b\theta_k = A(\theta_k - \bar{\theta}_k)^2 - A\bar{\theta}_k^2$, where $\bar{\theta}_k = A^{-1}b = (||\mathbf{d}_k||^2 + \sigma_{11.2}/\tau_{\theta_k}^2)^{-1}\mathbf{y}^{*T}\mathbf{d}_k$. We can move the latter term outside the integral, so that (4.12) becomes

$$g_2(\theta_k) = s_k c^* exp\left(\frac{A\bar{\theta}_k^2}{2\sigma_{11.2}}\right) \int exp\left(-\frac{A}{2\sigma_{11.2}}(\theta_k - \bar{\theta}_k)^2\right) d\theta_k \tag{4.13}$$

Notice that the integrand is now in the form of a Normal pdf, scaled by a constant. Thus, after integrating, we have

$$g_2(\theta_k) = s_k c^* exp\left(\frac{A\bar{\theta}_k^2}{2\sigma_{11.2}}\right)\left(\frac{2\pi\sigma_{11.2}}{A}\right)^{\frac{1}{2}}. \tag{4.14}$$

After solving both integral expressions, we can now write a proportional form of $p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k})$ explicitly. Since from (4.9) $p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto p(s_k)g(\theta_k)$ and $g(\theta_k) = g_1(\theta_k) + g_2(\theta_k)$, we can drop any constant terms which are common to both $g_1(\theta_k)$ and $g_2(\theta_k)$. Writing the likelihood function in (4.10) explicitly, we have

$$f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}_{-k}, \boldsymbol{\eta}, \theta_k = 0) = \frac{1}{(2\pi\sigma_{11.2})^{\frac{n}{2}}} exp\left(\frac{-1}{2\sigma_{11.2}}||\mathbf{y}^*||^2\right) = (2\pi\tau_{\theta_k}^2)^{\frac{1}{2}} c^*$$

which are terms contained in $g_2(\theta_k)$. Furthermore, the remaining $(2\pi\tau_{\theta_k}^2)^{-\frac{1}{2}}$ term from c in $g_2(\theta_k)$ can be combined with the $(2\pi\sigma_{11.2}/A)^{\frac{1}{2}}$ term, cancelling the 2π. After rearranging, the posterior distribution becomes proportional to

$$p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto p(s_k)\left[(1 - s_k) + s_k exp\left(\frac{A\bar{\theta}_k^2}{2\sigma_{11.2}}\right)\left(\frac{\frac{\sigma_{11.2}}{\tau_{\theta_k}^2}}{A}\right)^{\frac{1}{2}}\right]. \tag{4.15}$$

To simplify notation, we will define

$$Q = exp\left(\frac{A\bar{\theta}_k^2}{2\sigma_{11.2}}\right)\left(\frac{\frac{\sigma_{11.2}}{\tau_{\theta_k}^2}}{A}\right)^{\frac{1}{2}}$$

so that the proportional form of the posterior distribution can be written simply as $p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto \omega^{s_k}(1-\omega)^{1-s_k}(1-s_k+Qs_k)$. Notice that when $s_k = 1$, $p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto \omega Q$ and when $s_k = 0$, $p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \propto 1-\omega$. This implies that $p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k})$ follows a Bernoulli distribution with success probability $\omega Q[1-\omega+\omega Q]^{-1}$. Therefore,

$$p(s_k|\boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}) \sim Ber\left(\frac{\omega Q}{(1-\omega)+\omega Q}\right). \tag{4.16}$$

Here, we point out that the success probability of the Bernoulli distribution will always be less than 1, since $\omega \in (0,1)$. Thus, this sampling method will avoid the dimension-changing problem discussed earlier with the standard construction of the Gibbs sampler.

4.2.2 The Posterior Distribution of θ_k

Next, we derive the posterior of θ_k given $\boldsymbol{\theta}_{-k}$ and $\mathbf{s}$. To find the posterior distribution of θ_k, we have $p(\theta_k|\mathbf{y}, \boldsymbol{\theta}_{-k}, \mathbf{s}) \propto f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta})p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{s})$. Rewriting the prior of θ_k explicitly, this formulation becomes

$$p(\theta_k|\mathbf{y}, \boldsymbol{\theta}_{-k}, \mathbf{s}) \propto f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta})\left[(1-s_k)\delta_0 + s_k\phi(\theta_k|0, \tau_{\theta_k}^2)\right] \tag{4.17}$$

In this form, it is clear that we have two cases, either $s_k = 1$ or $s_k = 0$. For $s_k = 0$, the posterior distribution is proportional to the unit impulse function at 0 since $f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{s})\delta_0 \propto \delta_0$. This relation holds even though the likelihood function contains θ_k. This is because the likelihood function is in the form of a multivariate Normal density function which is always positive. This property is important, since it shows our method satisfies the sparsity condition to identify valid instrumental variables.

Next, we examine the case when $s_k = 1$. In the derivation of the full conditional distribution of s_k, the expression $f(\mathbf{y}|\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) s_k \phi(\theta_k|0, \tau^2_{\theta_k})$ is equal to the integrand of $g_2(\theta_k)$ given in (4.11). However, in this case we can drop the constant terms from the proportional form, since $s_k = 1$ is given in this case. Thus,

$$p(\theta_k|\mathbf{y}, \boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}, s_k = 1) \propto exp\left(-\frac{1}{2\sigma_{11.2}}||\mathbf{y}^* - \mathbf{d}_k\theta_k||^2 + \frac{\sigma_{11.2}}{\tau^2_{\theta_k}}\theta_k^2\right). \tag{4.18}$$

As before, we will write the expression in centered form, from (4.13) to obtain a proportional form of a Normal density function. We again define $\bar{\theta}_k = (||\mathbf{d}_k||^2 + \sigma_{11.2}/\tau^2_{\theta_k})^{-1}\mathbf{d}_k^T\mathbf{y}^*$. Then, the posterior distribution is proportional to a Normal distribution with mean $\bar{\theta}_k$ and variance $\sigma_{11.2}(||\mathbf{d}_k||^2 + \sigma_{11.2}/\tau^2_{\theta_k})^{-1}$:

$$p(\theta_k|\mathbf{y}, \boldsymbol{\theta}_{-k}, \mathbf{s}_{-k}, s_k = 1) \propto exp\left(-\frac{1}{2\sigma_{11.2}}\left(||\mathbf{d}_k||^2 + \frac{\sigma_{11.2}}{\tau^2_{\theta_k}}\right)(\theta_k - \bar{\theta}_k)^2\right) \tag{4.19}$$

To summarize, the full conditional distribution of θ_k follows the unit impulse function if $s_k = 0$ and follows a Normal distribution with mean $\bar{\theta}_k$ and variance $\sigma_{11.2}(||\mathbf{d}_k||^2 + \sigma_{11.2}/\tau^2_{\theta_k})^{-1}$ if $s_k = 1$. Note that for $\theta_0 = \beta_0$ and $\theta_1 = \beta_1$, we have defined $s_0 = s_1 = 1$. Thus, the full conditional distributions of β_0 and β_1 are Normal.

4.3 The Posterior Distribution of $\boldsymbol{\Sigma}$

Recall from (2.3) that the prior of $\boldsymbol{\Sigma}$ follows an inverse-Wishart distribution with scale matrix $\boldsymbol{\Psi}$ and degrees of freedom ν. That is,

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\nu+2+1}{2}} exp\left(-\frac{1}{2}tr(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right) \tag{4.20}$$

where $tr(\cdot)$ is the trace of a square matrix. Also, define $\boldsymbol{\xi}_i = (\epsilon_i, \eta_i)^T = (y_i - \mathbf{d}_i^T\boldsymbol{\theta}, x_i -$

$\mathbf{z}_i^T\boldsymbol{\alpha})^T$. We can then express the posterior distribution of $\boldsymbol{\Sigma}$ as the product of its prior with $f(\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\Sigma})$, the likelihood of $\boldsymbol{\xi}$. Furthermore, since $\boldsymbol{\xi}_i|\boldsymbol{\Sigma} \overset{iid}{\sim} N_2(0,\boldsymbol{\Sigma})$ we can express the likelihood as the product of multivariate normal densities. Thus,

$$p(\boldsymbol{\Sigma}|\mathbf{D},\boldsymbol{\alpha},\boldsymbol{\theta}) \propto \prod_{i=1}^{n} f(\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}) \tag{4.21}$$

Both the likelihood of $\boldsymbol{\xi}$ and the prior of $\boldsymbol{\Sigma}$ contain powers of $|\boldsymbol{\Sigma}|$. There are n products of $|\boldsymbol{\Sigma}|^{-\frac{1}{2}}$ from the likelihood, times $|\boldsymbol{\Sigma}|^{-\frac{\nu+3}{2}}$ from the prior distribution. Thus, this term becomes $|\boldsymbol{\Sigma}|^{-\frac{n+\nu+3}{2}}$. Now, the remaining portion of (4.21) are the exponential terms. The exponential terms from the likelihood simplify to $exp(\sum_{i=1}^{n}\boldsymbol{\xi}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}_i)$. Additionally, we can express the sum of this quadratic form using the trace function, since $\sum_{i=1}^{n}\boldsymbol{\xi}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}_i = \sum_{i=1}^{n} tr(\boldsymbol{\xi}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}_i)$. We can also swap the order of multiplication inside the trace, so that the sum becomes $\sum_{i=1}^{n} tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)$. Then, we bring the summation inside the trace function, so that $tr(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)$ can be easily merged with $tr(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}) = tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})$, the trace from the prior distribution. After combining and rearranging the terms as just described, the posterior distribution becomes proportional to

$$p(\boldsymbol{\Sigma}|\mathbf{D},\boldsymbol{\alpha},\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}|^{-\frac{n+\nu+3}{2}} exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \boldsymbol{\Psi}\right)\right]\right\}, \tag{4.22}$$

which is proportional to an inverse-Wishart distribution. Therefore, the posterior distribution of $\boldsymbol{\Sigma}$ is an inverse-Wishart distribution with $n+\nu$ degrees of freedom and scale matrix $\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \boldsymbol{\Psi}$, where $\boldsymbol{\xi}_i = (y_i - \mathbf{d}_i^T\boldsymbol{\theta}, x_i - \mathbf{z}_i^T\boldsymbol{\alpha})^T$ and $\mathbf{d}_i^T$ is the i^{th} row of $\mathbf{D}$.

Chapter 5

Simulation studies

To test the performance of our estimator, we performed several simulation studies and computed the MSE of several estimators. Our proposed estimator, referred to in Table 5.2 as Bayes UIIV, displays oracle properties in all cases, including when the proportion of invalid IVs approaches (but does not exceed) 50%.

Additionally, Table 5.2 displays the MSE for the Naive OLS, Oracle OLS, and Bayes Oracle estimators. Naive OLS refers to the standard ordinary least squares estimator, while Oracle OLS refers to the modification to the Two-Stage Least Square Estimator when the invalid IVs are known. The Bayes Oracle estimator is from a simplified form of the Gibbs sampler when $\boldsymbol{\gamma}$ is known. As shown in Table 5.2, our estimator demonstrates oracle properties in each of the simulation studies.

For each simulation study, we set the true value of $\beta_0 = 1, \beta_1 = 1.5$ and $\alpha_0 = 0$. Additionally, we set $\omega = 0.5, \tau_\alpha^2 = \tau_\beta^2 = \tau_\gamma^2 = 1000, \nu = 1$, and $\boldsymbol{\Psi} = \mathbf{I}_2$. We summarize simulation-specific parameter values in Table 5.1. N is the sample size for a data set in each iteration and p is the size of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. $\boldsymbol{\alpha}$ is set to $\mathbf{1}$ times a constant, where $\mathbf{1}$ is a $p \times 1$ vector of 1's. L is an index set specifying the invalid IVs in the simulation, and q is the

Table 5.1: *Simulation Study Parameters*

Simulation	N	p	$\boldsymbol{\alpha}_1$	L	q
UIIV(1)	200	100	$\mathbf{1}$	(99,100)	2
UIIV(2)	200	100	$\mathbf{1}$	(91:100)	10
WIV(1)	200	100	$0.5 \cdot \mathbf{1}$	(99,100)	2
WIV(2)	200	100	$0.1 \cdot \mathbf{1}$	(99,100)	2
Many Invalid	500	100	$\mathbf{1}$	(52:100)	49

A summary of the parameters changed for each simulation study. N is the sample size, p is the dimension of the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_1$ is the strength of the instrumental variables where $\mathbf{1}$ is a $p \times 1$ vector of ones, L is an index set indicating the non-zero elements of $\boldsymbol{\gamma}$ corresponding to invalid instrumental variables, and q is the number of invalid instruments. The first two studies, UIIV(1) and UIIV(2) simply vary the number of invalid instruments, while WIV(1) and WIV(2) weaken the strength of the instrumental variables. The Many Invalid setting demonstrates the case where the number of invalid instruments is just below 50% of all instruments.

Table 5.2: *MSE of Estimators*

Estimator	UIIV(1)	UIIV(2)	WIV(1)	WIV(2)	Many Invalid
Naive OLS	0.270508	0.296745	0.269202	0.250123	0.248497
Oracle OLS	0.000062	0.000043	0.000434	0.030910	0.000050
Bayes UIIV	0.000059	0.000053	0.000345	0.029203	0.000041
Bayes Oracle	0.000060	0.000052	0.000309	0.029067	0.000017

The estimated MSE of each estimator for each simulation study, based on 100 simulated data sets. The Naive OLS is standard ordinary least squares regression while Oracle OLS is the generalized method of moments modification to the OLS estimator when the set of invalid IVs are known. Bayes UIIV is the estimator proposed in this paper and Bayes Oracle is the same Gibbs sampler but with known $\boldsymbol{\gamma}$.

number of invalid instruments. We also set all invalid $\gamma_j = 1$. In Table 5.2, we summarize the MSE of several estimators. The MSE is computed based on 100 simulated data sets for each simulation study.

Chapter 6

Real data analysis

For our real data analysis, we used the trade share dataset discussed in Fan and Wu (2022)[5]. We used their R package, naivereg, to obtain the dataset. The dataset consists of 158 countries with 20 columns consisting of the response variable, endogenous variable of interest, and 18 potential instruments. The response variable is a country's GDP per worker (in log units) and the endogenous variable is the "share of international trade to GDP" (Fan and Wu 2022)[5]. The instruments used are the constructed trade share, population size (in log units), land area (in log units), water area, total length of coastline, percentage of arable land, total length of land borders, percentage of forested land, the number of official languages, and the air quality index PM2.5. Additionally, the interactions of constructed trade share with water area, coastline, arable land percentage, border length, forest percentage, and number of languages are included as potential instruments.

Following Fan and Wu (2022)[5], we modeled the data as

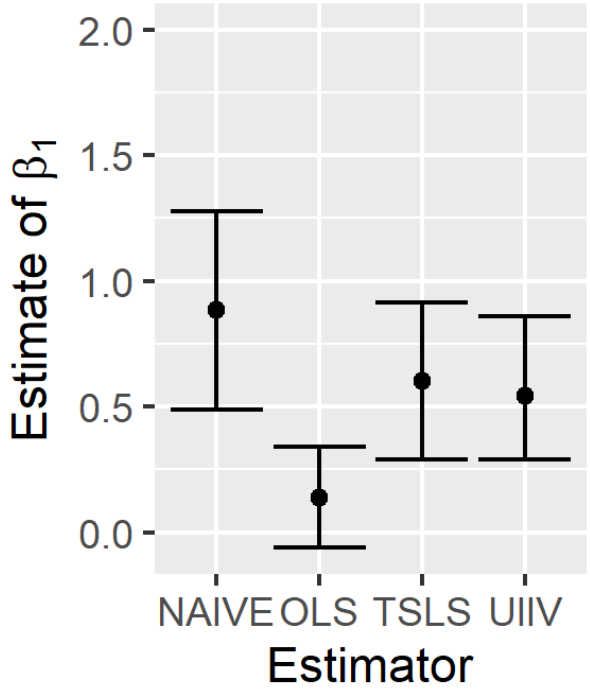
$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \mathbf{z}_i^T \boldsymbol{\gamma} + \epsilon_i \\ x_i &= \alpha_0 + \mathbf{z}_i^T \boldsymbol{\alpha} + \eta_i \end{aligned} \tag{6.1}$$

where y_i is log income per capita for country i, x_i is real trade share (the endogenous

variable of interest), and $\mathbf{z}_i^T$ are all other predictors after regressing on the spatial variables land area and population (in log units). We include the quadratic and cubic terms of each of the instrumental variables in $\mathbf{z}_i^T$ to model a nonlinear relationship between $\mathbf{z}_i^T$ and x_i as described in Fan and Wu (2022)[5]. We also generated two variables from standard normal distributions to include in $\mathbf{z}_i^T$ as in Fan and Wu (2022)[5].

Under this framework, our estimator identified the atmospheric pollution variable, PM2.5, as the only invalid instrumental variable (including its squared and cubic terms), which is consistent with the results of R2IVE in Fan and Wu (2022)[5]. As shown in Figure 6.1, our estimate of the endogenous effect was 0.56, with a 95% credible interval of (0.299, 0.837). The sample size was 158, and we set $\omega = 0.5, \tau_\alpha^2 = \tau_\beta^2 = \tau_\gamma^2 = 1000, \nu = 1$, and $\mathbf{\Psi} = \mathbf{I}_2$. A comparison of the results from several estimators is given in Figure 6.1.

Figure 6.1: *Trade Share Estimation Results*



A plot of the estimates of β_1, the effect of global trade share on GDP per worker, with 95% confidence bars. UIIV is the proposed method and the band represent a credible interval instead of a confidence interval. NAIVE is an estimator used in Fan and Wu (2022)[5], included for reference. OLS is the standard ordinary least square estimator and TSLS is the two stage least square estimator.

Chapter 7

Concluding remarks

We have demonstrated that our proposed estimator identifies invalid instrumental variables and estimates the endogenous effect with oracle properties. Furthermore, our estimator has the advantage of removing model uncertainty by integrating over the prior distributions of the parameters, in contrast to penalized regression approaches which require proper selection of a tuning parameter. Our choice of prior distributions also constrain the applicability of our method, however, the framework presented in this paper can be easily extended to a variety of settings.

For example, the model we presented assumed normally distributed priors, which led to a normal posterior distribution for β_1. For applications which account for extreme events, a heavier-tailed distribution may be more appropriate. Extending our estimator to include distributions with heavier tails can be easily done by adding a layer to the hierarchical model. For example, β_1 will have a posterior t-distribution if we assign a Normal prior to β_1 conditional on τ_β^2 and assign an inverse-Gamma prior to τ_β^2.

Additionally, our paper focuses on the identification of invalid instrumental variables from a set of known instruments. That is, we assume $\alpha_j \neq 0$ for all j. Other approaches,

such as R2IVE in Fan and Wu (2022)[5] differ by allowing $\alpha_j = 0$ for some j, which are termed irrelevant instruments. However, our method can be extended to estimate sparse $\boldsymbol{\alpha}$ by modifying its prior distribution to be conditional on a set of Bernoulli random variables, similar to the set up for $\boldsymbol{\gamma}$. Then, after appropriately adjusting the Gibbs sampler, we could identify irrelevant instruments in addition to valid and invalid instruments, similar to the framework described in Fan and Wu (2022)[5]. We leave these extensions for further research.

Bibliography

[1] Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

[2] Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019. PMID: 31708716.

[3] Gyuhyeong Goh and Jisang Yu. Causal inference with some invalid instrumental variables: A quasi-bayesian approach*. *Oxford Bulletin of Economics and Statistics*, 84(6): 1432–1451, 2022.

[4] Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[5] Qingliang Fan and Yaqian Wu. Endogenous Treatment Effect Estimation with a Large and Mixed Set of Instruments and Control Variables. *The Review of Economics and Statistics*, pages 1–45, 07 2022.