MSCI Project Report

Bayesian Analysis of Linear Models with Endogeneity using Instrumental Variables

Name: Dongyuan Lin
Student Number: 530814230
Supervisor: Associate Professor Clara Grazian & Dr Linh Nghiem

Contents

Chapter 1

Introduction

1.1 Background and Motivation

In real-world applications, data analysis frequently faces two major challenges: endogeneity and the difficulty of variable selection in sparse setting contexts.

Endogeneity arises when explanatory variables are correlated with the error term, violating a core assumption of the ordinary least squares (OLS) method. This problem is especially prevalent in observational studies, where omitted variable bias, measurement error, or reverse causality can easily occur. Failing to account for endogeneity may result in biased and inconsistent estimates, leading to flawed causal inference and misguided policy recommendations.

At the same time, modern data environments often involve a large number of potential predictors—sometimes exceeding the number of observations—posing a serious challenge for variable selection. Identifying the truly relevant covariates is crucial not only for interpretability and predictive accuracy but also for addressing computational issues associated with sparse setting.

In practice, these two issues—endogeneity and high-dimensional variable selection—frequently coexist, necessitating a unified framework that can effectively address both. The Bayesian approach offers a natural and flexible solution. By incorporating prior information and performing full posterior inference, Bayesian methods provide a coherent framework for handling model uncertainty, integrating variable selection, and dealing with endogenous regressors. This report explores the application of Bayesian methods to these dual challenges, particularly through instrumental variable techniques and variable selection strategies.

1.2 Endogeneity Problem in Regression Models

One of the key assumptions in the standard linear regression model is that the explanatory variables are uncorrelated with the error term. When this assumption is violated, the model suffers from the problem of endogeneity. Formally, consider the linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is the outcome variable, $\mathbf{X}$ is the matrix of explanatory variables, $\boldsymbol{\beta}$ is the vector of coefficients, and $\boldsymbol{\varepsilon}$ is the error term. If

$$\mathbb{E}[\mathbf{X}^\top \boldsymbol{\varepsilon}] \neq \mathbf{0},$$

then $\mathbf{X}$ is said to be endogenous, and ordinary least squares (OLS) estimators become biased and inconsistent.

Endogeneity can arise from a variety of sources (Hill et al., 2021):

- **Omitted variable bias**: When relevant variables that affect both the explanatory and outcome variables are left out of the model, the included regressors become correlated with the error term.
- **Measurement error**: Errors in measuring the explanatory variables lead to discrepancies between observed and true values, inducing correlation between the measured regressors and the error term. This is especially common in survey-based or self-reported data.
- **Simultaneity or reverse causality**: When the outcome variable also influences the explanatory variable, a two-way causality arises.
- **Sample selection bias**: If the process of sample selection is related to both the regressors and the outcome, the sample becomes unrepresentative of the population, resulting in biased estimation.

Endogeneity poses a serious threat to causal inference. Biased coefficient estimates can lead to incorrect conclusions about both the magnitude and direction of causal effects, which in turn may result in flawed policy recommendations, ineffective interventions, or misleading scientific theories.

A classical remedy for endogeneity is the instrumental variables (IV) approach, which introduces instruments—variables that are correlated with the endogenous regressors but uncorrelated with the error term—to obtain consistent estimates of causal effects. The basic IV model can be expressed as the following system of equations:

$$\begin{aligned} \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma} + \boldsymbol{U}, \\ \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \end{aligned} \tag{1.2.1}$$

where $\mathbf{Z}$ is a matrix of instrumental variables, $\boldsymbol{\Gamma}$ captures the relationship between instruments and endogenous regressors, and $(\boldsymbol{U}, \boldsymbol{\varepsilon})$ are error terms that may be correlated with each other but are assumed to be uncorrelated with $\mathbf{Z}$.

Traditional frequentist IV estimation methods, such as two-stage least squares (2SLS), provide consistent point estimates but face limitations in small samples, in handling weak instruments, and in incorporating prior information. Bayesian methods offer a promising alternative that addresses these shortcomings while retaining the ability to correct for endogeneity.

1.3 Variable Selection Problem

The variable selection problem can be framed as the task of identifying a subset of predictors that strikes an optimal balance between model fit and model complexity. In the context of linear regression, this involves selecting a subset of columns from the design matrix $\mathbf{X}$ that best explains the response variable $\mathbf{Y}$, while avoiding overfitting.

Several challenges make variable selection nontrivial (Wasserman and Roeder, 2009):

- **Computational complexity**: The number of possible models grows exponentially with the number of predictors, making exhaustive search infeasible even for problems of moderate size.
- **Multicollinearity**: When predictors are highly correlated, traditional selection methods may arbitrarily select among them, leading to model instability.
- **Bias-variance trade-off**: Including too few predictors may introduce bias (underfitting), whereas including too many can increase variance (overfitting).
- **Lack of uncertainty quantification**: Traditional variable selection methods typically produce a single "best" model without quantifying the uncertainty associated with the selection process.

Various methods have been developed to address some of the above challenges, particularly those related to model complexity, multicollinearity, and the bias-variance trade-off. Classical approaches include stepwise selection procedures (forward, backward, or bidirectional), information criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), as well as regularization techniques like ridge regression and the LASSO.

However, these methods often focus on producing a single "best" model and typically lack formal mechanisms for quantifying uncertainty in the selection process, which remains an important limitation. Bayesian approaches, by contrast, offer a coherent probabilistic framework for addressing both variable selection and uncertainty quantification simultaneously.

The LASSO has gained widespread popularity due to its ability to perform parameter estimation and variable selection (Lin et al., 2015). It achieves this by adding an ℓ_1 penalty term to the regression objective:

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is a regularization parameter controlling the strength of the penalty (Park and Casella, 2008). The ℓ_1 penalty encourages sparsity in the coefficient vector, leading some coefficients to be exactly zero, and thus effectively selecting a subset of predictors.

While LASSO and related techniques offer effective solutions to the variable selection problem in standard regression settings, they face additional challenges in the presence of endogeneity. Correlation between predictors and the error term can distort the selection process, resulting in the inclusion of variables that appear important due to their correlation with the error. This highlights the need for methods capable of addressing endogeneity and variable selection jointly.

Bayesian approaches provide a natural framework for variable selection by treating model structure as part of the inferential process. These methods allow the incorporation of prior beliefs about the relevance of predictors, quantify uncertainty about which variables should be included, and can seamlessly integrate with strategies for correcting endogeneity.

1.4 Bayesian Approaches

The central idea of Bayesian inference lies in updating prior beliefs about model parameters using the likelihood of the observed data to produce a posterior distribution, which reflects our updated knowledge after seeing the data. Despite its potential computational complexity, the joint posterior distribution encodes all available information about the model parameters by combining the prior and the likelihood. For linear regression models, the Bayesian framework can be expressed as:

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X}) \propto p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2),$$

where $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X})$ is the posterior distribution of the regression coefficients $\boldsymbol{\beta}$ and the error variance σ^2, $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2)$ is the likelihood function, and $p(\boldsymbol{\beta}, \sigma^2)$ is the prior distribution.

In the standard Bayesian linear regression model (denoted as Naive), which does not account for endogeneity, the likelihood is typically specified as:

$$\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

This model assumes homoscedasticity, independence of errors, and that the covariates are exogenous.

Priors are usually chosen for computational convenience or to reflect substantive prior knowledge:

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_\beta, \sigma_\beta^2\mathbf{I}),$$

$$\sigma^2 \sim \text{Inverse-Gamma}(a, b).$$

This simple model assumes that the predictors $\mathbf{X}$ are exogenous. However, when endogeneity is present, more sophisticated models are required to account for the correlation between the regressors and the error term.

The Bayesian instrumental variables (IV) approach extends this framework by modeling the joint distribution of the outcome, endogenous regressors, and instruments. The basic Bayesian IV model is represented by the system (1.2.1). The error terms are jointly distributed as:

$$\begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_u & \boldsymbol{\Sigma}_{u\varepsilon} \\ \boldsymbol{\Sigma}_{u\varepsilon}^\top & \sigma_\varepsilon^2 \end{pmatrix} \right)$$

Naive serves as a foundation for the more advanced Bayesian IV models presented in subsequent sections. While Naive itself does not correct for endogeneity, it offers a conceptual starting point for understanding regression modeling under the Bayesian paradigm. The more advanced models that simultaneously address endogeneity and variable selection will be developed in later sections.

In the context of variable selection, Bayesian methods offer similarly appealing properties. The Bayesian LASSO extends the classical LASSO by interpreting the ℓ_1 penalty as a prior distribution. Specifically, the LASSO estimate can be viewed as the posterior mode under a Laplace (double-exponential) prior on each regression coefficient:

$$p(\beta_j \mid \sigma^2) = \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}} \right)$$

This prior induces sparsity in the posterior distribution, encouraging some coefficients to shrink exactly to zero. Unlike the frequentist LASSO, which yields a single sparse solution, the Bayesian LASSO provides a full posterior distribution for the coefficients, allowing uncertainty to be quantified during the variable selection process.

By combining the Bayesian IV approach with Bayesian variable selection techniques, researchers can simultaneously address endogeneity and sparse setting within a unified framework. The subsequent sections of this paper will elaborate on these methods, with particular attention to the Bayesian LASSO in the context of instrumental variable regression.

1.5 MCMC Implementation

To conduct posterior inference, we adopt a Markov Chain Monte Carlo (MCMC) framework, with a particular focus on Gibbs sampling. This method generates samples by iteratively drawing from the full conditional distributions of each model parameter, conditioning on the current values of all other parameters in the system. When these full conditionals belong to standard distribution families—such as the multivariate normal or inverse-Wishart—the sampling procedure becomes both straightforward to implement and computationally efficient.

Over successive iterations, the Gibbs sampler produces a sequence of dependent draws that, under regularity conditions, converge in distribution to the joint posterior of the parameters. These posterior samples can then be used to compute summary statistics such as means, variances, and quantiles, and to construct credible intervals for the parameters of interest.

1.6 Thesis Structure and Notation Guide

Chapter 1 provides an introduction to two pervasive challenges in statistical modeling—endogeneity and variable selection—and discusses why the Bayesian framework offers a promising approach to addressing both issues.

Chapter 2 presents our Bayesian instrumental variable models in detail, including the full Bayesian method and the simplified Bayesian approach. Extensive simulation studies demonstrate that the simplified Bayesian outperforms the full Bayesian across a range of scenarios. This chapter also includes an application of these methods to real-world data.

Chapter 3 incorporates the Bayesian LASSO into our framework to address variable selection within instrumental variable models. We show how this approach enables effective selection of relevant predictors while retaining the ability to correct for endogeneity. The advantages of this method are illustrated through simulation experiments and empirical examples.

Chapter 4 summarizes the main findings, discusses the limitations of the proposed methods, and outlines directions for future research.

We use the following notations throughout the thesis. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we let $\text{tr}(\mathbf{A})$ be the trace, i.e. sum of all the diagonal elements of $\mathbf{A}$, $\text{vec}(\mathbf{A})$ be the vectorization operator. We use $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$.

Similarly, $\mathcal{MN}_{a\times b}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ denotes the matrix normal distribution over $a \times b$ random matrices, with mean matrix $\mathbf{M} \in \mathbb{R}^{a\times b}$, row covariance $\mathbf{U} \in \mathbb{R}^{a\times a}$, and column covariance $\mathbf{V} \in \mathbb{R}^{b\times b}$; the subscript $a \times b$ indicates the dimension of the random matrix. We use $p(\cdot)$ to denote a generic probability density function, with its argument specifying the distributional variable. Finally, AR(1) refers to a first-order autoregressive process defined as $x_t = \rho x_{t-1} + \varepsilon_t$, where $|\rho| < 1$ and ε_t is white noise.

Chapter 2

Bayesian Instrumental Variables Model

2.1 Model Specification

2.1.1 Basic Instrumental Variables Model

The model is specified in Equation (1.2.1), where $i = 1, 2, \ldots, n$ indexes individual observations. The matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ represents the endogenous regressors, with each row vector $\mathbf{X}_i \in \mathbb{R}^{1 \times p}$ corresponding to the i-th observation. Instrumental variables are collected in the matrix $\mathbf{Z} \in \mathbb{R}^{n \times q}$, where $\mathbf{Z}_i \in \mathbb{R}^{1 \times q}$ denotes the instrument vector for the i-th unit. The relationship between instruments and endogenous variables is governed by the coefficient matrix $\mathbf{\Gamma} \in \mathbb{R}^{q \times p}$, and the associated first-stage disturbances are captured in the error matrix $\mathbf{U} \in \mathbb{R}^{n \times p}$. The outcome variable is recorded in the vector $\mathbf{Y} \in \mathbb{R}^{n \times 1}$, with each scalar Y_i representing the response for observation i. The vector of second-stage regression coefficients is denoted by $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$, and the corresponding second-stage error terms are contained in the vector $\boldsymbol{\varepsilon} \in \mathbb{R}^{n \times 1}$.

2.1.2 Statistical Assumptions

In the instrumental variable (IV) model, we make several key assumptions regarding the error terms:

- **Independence across observations.** It is assumed that the error terms are independent across different observations. That is, for any $i \neq j$, the pairs $(\mathbf{U}_i, \varepsilon_i)$ and $(\mathbf{U}_j, \varepsilon_j)$ are mutually independent.
- **Identical distribution.** The error terms are assumed to follow the same distribution across observations. For each observation i, the joint distribution of the errors is:
$$\begin{bmatrix} \mathbf{U}_i \\ \varepsilon_i \end{bmatrix} \sim \mathcal{N}_{p+1} \left(\begin{bmatrix} \mathbf{0}_p \\ 0 \end{bmatrix}, \mathbf{\Sigma} \right),$$
where $\mathbf{\Sigma} \in \mathbb{R}^{(p+1) \times (p+1)}$ is a covariance matrix defined as:
$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_u & \mathbf{\Sigma}_{u\varepsilon} \\ \mathbf{\Sigma}_{u\varepsilon}^\top & \sigma_\varepsilon^2 \end{bmatrix}, \tag{2.1.1}$$

with $\boldsymbol{\Sigma}_u \in \mathbb{R}^{p\times p}$ being the covariance matrix of the error term $\mathbf{U}_i$; $\boldsymbol{\Sigma}_{u\varepsilon} \in \mathbb{R}^{p\times 1}$ denoting the vector of covariances between $\mathbf{U}_i$ and ε_i; and $\sigma_\varepsilon^2 \in \mathbb{R}$ representing the variance of the structural error ε_i.

- **Endogeneity structure.** If $\boldsymbol{\Sigma}_{u\varepsilon} \neq \mathbf{0}$, the errors in the endogenous regressors $\mathbf{U}_i$ are correlated with the structural error ε_i, which is the defining feature of endogeneity.

- **Instrument validity.** It is assumed that the instruments are uncorrelated with the error terms:

$$\mathbb{E}[\mathbf{Z}^\top \mathbf{U}] = \mathbf{0}, \quad \mathbb{E}[\mathbf{Z}^\top \varepsilon] = \mathbf{0}.$$

  This ensures that the instruments affect the outcome only through their relationship with the endogenous regressors.

- **Instrument relevance.** The instruments must be sufficiently correlated with the endogenous variables to ensure model identification. Specifically, the matrix $\mathbf{Z}^\top \mathbf{X}$ must have full rank, which typically requires at least as many valid instruments as endogenous regressors.

These assumptions collectively ensure that consistent estimation of the causal parameter $\boldsymbol{\beta}$ is possible through the IV approach, even when endogeneity is present.

2.1.3 Prior Distribution

In the Bayesian framework, it is necessary to specify prior distributions for all model parameters. We specify proper priors that guarantee the resulting posterior distribution is also proper. We adopt the following prior specifications:

- **Second-stage coefficients $\boldsymbol{\beta}$:**

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_\beta, \sigma_\beta^2 \mathbf{I}_p),$$

  where $\boldsymbol{\mu}_\beta$ is the prior mean vector, typically set to a zero vector, and σ_β^2 is the prior variance, often chosen to be large to reflect weakly informative prior beliefs.

- **First-stage coefficients $\boldsymbol{\Gamma}$:**

$$\boldsymbol{\Gamma} \sim \mathcal{MN}_{q\times p}(\boldsymbol{\mu}_\Gamma, \mathbf{I}_q, \sigma_\Gamma^2 \mathbf{I}_p),$$

  where $\boldsymbol{\mu}_\Gamma$ is the prior mean matrix, usually set to a zero matrix, and σ_Γ^2 is the prior variance, also commonly set to a large value to reflect vague prior information.

- **Covariance matrix $\boldsymbol{\Sigma}$:**

$$\boldsymbol{\Sigma} \sim \text{Inverse-Wishart}_{p+1}(\nu_0, \boldsymbol{\Psi}_0),$$

  where ν_0 is the degrees of freedom and $\boldsymbol{\Psi}_0$ is the scale matrix. This prior ensures that sampled covariance matrices are positive definite.

These priors are conjugate, which facilitates analytical tractability and simplifies posterior computations. In particular, the full conditional distributions for all parameters admit closed-form expressions, making them directly sampleable via Gibbs sampling. In practice, we typically employ non-informative or weakly informative priors so that the data dominate the posterior inference. However, when reliable prior knowledge is available, these distributions can be modified accordingly to reflect that information.

2.1.4 Posterior Distribution

In the Bayesian framework, inference is based on the posterior distribution, which combines prior beliefs about the parameters with the information contained in the observed data. For the instrumental variables model, the likelihood function depends on both stages of the structural system, and the full joint posterior includes the second-stage coefficients, the first-stage coefficients, and the joint error covariance matrix. The full conditional distributions are tractable under conjugate priors, allowing for efficient Gibbs sampling.

Bayesian analysis proceeds by drawing samples from the joint posterior distribution, from which marginal posterior distributions for individual parameters can be easily obtained. Since this is analytically intractable for most models of practical interest, we rely on sampling-based methods to approximate the posterior distribution.

2.2 Full Bayesian Approach

We now derive the full conditional distributions for the parameters under full Bayesian approach.

2.2.1 Joint Likelihood Formulation

Full Bayesian addresses endogeneity by directly modeling the joint distribution of the outcome variable $\mathbf{Y}$ and the endogenous regressors $\mathbf{X}$. Rather than treating the two stages separately, this approach fully captures the dependence structure between the error terms.

Under our model specification and statistical assumptions, the joint distribution for each observation can be expressed as:

$$\begin{pmatrix} \mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma} \\ Y_i - \mathbf{X}_i\boldsymbol{\beta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0}_p \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} \right).$$

Assuming independence across observations, the joint likelihood of the full dataset is given by:

$$p(\mathbf{Y}, \mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) = (2\pi)^{-n(p+1)/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma} \\ Y_i - \mathbf{X}_i\boldsymbol{\beta} \end{pmatrix}^\top \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma} \\ Y_i - \mathbf{X}_i\boldsymbol{\beta} \end{pmatrix} \right\}.$$

Based on the block matrix inverse formula, for the covariance matrix (2.1.1), its inverse is:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_u^{-1} + \frac{1}{\sigma^2_{\varepsilon|u}} \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon} \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} & -\frac{1}{\sigma^2_{\varepsilon|u}} \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon} \\ -\frac{1}{\sigma^2_{\varepsilon|u}} \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} & \frac{1}{\sigma^2_{\varepsilon|u}} \end{pmatrix}, \tag{2.2.1}$$

where $\sigma^2_{\varepsilon|u} = \sigma^2_\varepsilon - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon}$ is the conditional variance. To facilitate the inversion of the joint covariance matrix $\boldsymbol{\Sigma}$, we introduce the conditional variance $\sigma^2_{\varepsilon|u}$, which corresponds to the Schur complement of $\boldsymbol{\Sigma}_u$ in $\boldsymbol{\Sigma}$. We note that $\sigma^2_{\varepsilon|u} > 0$ under the assumption that $\boldsymbol{\Sigma}$ is positive definite. Let $\mathbf{U}_i = \mathbf{X}_i - \mathbf{Z}_i \boldsymbol{\Gamma}$ and $\varepsilon_i = Y_i - \mathbf{X}_i \boldsymbol{\beta}$. Expanding and reorganizing the quadratic form, we obtain:

$$\begin{pmatrix} \mathbf{U}_i \\ \varepsilon_i \end{pmatrix}^\top \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \mathbf{U}_i \\ \varepsilon_i \end{pmatrix} = \mathbf{U}_i^T \boldsymbol{\Sigma}_u^{-1} \mathbf{U}_i + \frac{(\varepsilon_i - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \mathbf{U}_i)^2}{\sigma^2_{\varepsilon|u}}.$$

Summing over all observations and converting to matrix form, the joint likelihood can be decomposed as:

$$\begin{aligned} p(\mathbf{Y}, \mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto \exp\left\{ -\frac{1}{2} \mathrm{tr}\left((\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma})^\top \boldsymbol{\Sigma}_u^{-1} (\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma}) \right) \right\} \\ \times \exp\left\{ -\frac{1}{2\sigma^2_{\varepsilon|u}} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i \boldsymbol{\beta} - \boldsymbol{\Sigma}_{u\varepsilon}^\top \boldsymbol{\Sigma}_u^{-1} (\mathbf{X}_i - \mathbf{Z}_i \boldsymbol{\Gamma}) \right)^2 \right\}. \end{aligned} \tag{2.2.2}$$

This decomposition expresses the joint likelihood as the product of two parts: the first part, $p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_u)$, describes the instrumental variable model. The second part, $p(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$, corresponds to a conditional normal regression of $\mathbf{Y}$ on $\mathbf{X}$, where the mean is adjusted to account for the endogeneity arising from the dependence between $\mathbf{X}$ and the error term. This dependence is induced by a non-zero cross-covariance term $\boldsymbol{\Sigma}_{u\varepsilon}$, which captures the correlation between the first-stage and second-stage errors. As a result, endogeneity is properly handled through the conditional variance $\sigma^2_{\varepsilon|u}$.

This joint likelihood formulation explicitly captures the correlation between the first-stage error $\mathbf{U}$ and the second-stage error $\boldsymbol{\varepsilon}$, which is the source of endogeneity. By modeling this dependence directly, the full Bayesian approach enables consistent estimation of model parameters while properly accounting for endogeneity.

2.2.2 Joint Posterior Distribution

Based on the joint likelihood derived in the previous section and the prior distributions specified in Section 2.1.3, we can construct the joint posterior distribution of the model parameters. By Bayes' theorem, the posterior distribution is proportional to the product of the likelihood and the prior distributions:

$$p(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}) \propto p(\mathbf{Y}, \mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \times p(\boldsymbol{\beta}) \times p(\boldsymbol{\Gamma}) \times p(\boldsymbol{\Sigma}).$$

Substituting the likelihood (2.2.2) and prior expressions, we obtain

$$
\begin{aligned}
p(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}) \propto & |\boldsymbol{\Sigma}|^{-(n+\nu_0+p+2)/2} \\
& \times \exp\Bigg\{-\frac{1}{2}\Big[\operatorname{tr}\Big((\mathbf{X}-\mathbf{Z}\boldsymbol{\Gamma})^\top \boldsymbol{\Sigma}_u^{-1}(\mathbf{X}-\mathbf{Z}\boldsymbol{\Gamma})\Big) \\
& + \frac{1}{\sigma^2_{\varepsilon|u}} \sum_{i=1}^{n} \Big(Y_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{\Sigma}_{u\varepsilon}^\top \boldsymbol{\Sigma}_u^{-1}(\boldsymbol{X}_i - \boldsymbol{Z}_i\boldsymbol{\Gamma})\Big)^2 \\
& + \frac{1}{\sigma^2_\beta}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^\top(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta) \\
& + \frac{1}{\sigma^2_\Gamma}\operatorname{tr}\Big[(\boldsymbol{\Gamma}-\boldsymbol{\mu}_\Gamma)^\top(\boldsymbol{\Gamma}-\boldsymbol{\mu}_\Gamma)\Big] \\
& + \operatorname{tr}(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1})\Big]\Bigg\},
\end{aligned}
\tag{2.2.3}
$$

where $\sigma^2_{\varepsilon|u} = \sigma^2_\varepsilon - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon}$.

The joint posterior distribution is not available in closed form, which makes direct sampling intractable and motivates the use of MCMC methods. The expression reveals complex dependencies among parameters, particularly the interaction between $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and the covariance matrix $\boldsymbol{\Sigma}$.

Despite its computational complexity, the joint posterior contains all information about the model parameters, fully incorporating the correlation structure introduced by endogeneity. To draw inference and perform parameter estimation from this distribution, we employ Markov Chain Monte Carlo (MCMC) methods—specifically, Gibbs sampling—to iteratively sample from the full conditional distributions of each parameter.

In the next section, we derive these full conditionals, which serve as the basis for Gibbs sampling. By sampling from each parameter's full conditional distribution in turn, we can efficiently explore the joint posterior without needing to directly handle its full analytical form.

2.2.3 Full Conditional Distribution for β

To derive the full conditional distribution for $\boldsymbol{\beta}$, we extract all terms from the joint posterior distribution in Equation (2.2.3) that involve $\boldsymbol{\beta}$. This leads to the following expression:

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto & \exp\left\{-\frac{1}{2\sigma^2_{\varepsilon|u}} \sum_{i=1}^{n} \Big(Y_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1}(\mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma})^T\Big)^2\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma^2_\beta}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^T(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\right\}
\end{aligned}
$$

To simplify the likelihood term, we introduce the notation

$$
W_i = Y_i - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1}(\mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma})^T, \tag{2.2.4}
$$

which rewrites the first exponential term as

$$
\exp\left\{-\frac{1}{2\sigma^2_{\varepsilon|u}} \sum_{i=1}^{n} (W_i - \mathbf{X}_i\boldsymbol{\beta})^2\right\}.
$$

Next, we expand the square in the exponent:

$$(W_i - \mathbf{X}_i\boldsymbol{\beta})^2 = W_i^2 - 2W_i\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{X}_i\boldsymbol{\beta})^2.$$

Substituting this expansion back into the exponential, we obtain:

$$\exp\left\{-\frac{1}{2\sigma^2_{\varepsilon|u}}\sum_{i=1}^{n}\left[W_i^2 - 2W_i\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{X}_i\boldsymbol{\beta})^2\right]\right\}.$$

Noting that the term $\sum_{i=1}^{n} W_i^2$ does not depend on $\boldsymbol{\beta}$, we drop it as a constant in the proportional expression. This yields:

$$\exp\left\{-\frac{1}{2\sigma^2_{\varepsilon|u}}\left[-2\sum_{i=1}^{n} W_i\mathbf{X}_i\boldsymbol{\beta} + \sum_{i=1}^{n}(\mathbf{X}_i\boldsymbol{\beta})^2\right]\right\}.$$

Rearranging the terms in the exponent, the expression becomes:

$$\exp\left\{-\frac{1}{2\sigma^2_{\varepsilon|u}}\sum_{i=1}^{n}(\mathbf{X}_i\boldsymbol{\beta})^2 + \frac{1}{\sigma^2_{\varepsilon|u}}\sum_{i=1}^{n} W_i\mathbf{X}_i\boldsymbol{\beta}\right\}.$$

We now turn to the prior term and expand its quadratic form as follows:

$$\begin{aligned}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)^T(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) &= \boldsymbol{\beta}^T\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\beta^T\boldsymbol{\beta} + \boldsymbol{\mu}_\beta^T\boldsymbol{\mu}_\beta \\ &= \boldsymbol{\beta}^T\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^T\boldsymbol{\mu}_\beta,\end{aligned}$$

where we have used the fact that $\boldsymbol{\beta}^T\boldsymbol{\mu}_\beta$ is a scalar and therefore equal to its transpose.

Combining the likelihood and prior components and dropping constants independent of $\boldsymbol{\beta}$, we collect all remaining terms into a single exponent:

$$\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2_{\varepsilon|u}}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \frac{1}{\sigma^2_\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\mu}_\beta) - \frac{2}{\sigma^2_{\varepsilon|u}}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\right]\right\}.$$

This expression can be reorganized as:

$$\exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^T\left(\frac{1}{\sigma^2_{\varepsilon|u}}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma^2_\beta}\mathbf{I}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\left(\frac{1}{\sigma^2_{\varepsilon|u}}\mathbf{X}^T\mathbf{W} + \frac{1}{\sigma^2_\beta}\boldsymbol{\mu}_\beta\right)\right]\right\}.$$

To simplify the notation, we define the following expressions for the posterior precision matrix and mean:

$$\tilde{\boldsymbol{\Sigma}}_\beta^{-1} = \frac{1}{\sigma^2_{\varepsilon|u}}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma^2_\beta}\mathbf{I}, \quad \tilde{\boldsymbol{\mu}}_\beta = \tilde{\boldsymbol{\Sigma}}_\beta\left(\frac{1}{\sigma^2_{\varepsilon|u}}\mathbf{X}^T\mathbf{W} + \frac{1}{\sigma^2_\beta}\boldsymbol{\mu}_\beta\right). \tag{2.2.5}$$

The exponent can now be recognized as the quadratic form of a multivariate normal kernel:

$$\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_\beta^{-1}\tilde{\boldsymbol{\mu}}_\beta = (\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}_\beta)^T\tilde{\boldsymbol{\Sigma}}_\beta^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}_\beta) - \tilde{\boldsymbol{\mu}}_\beta^T\tilde{\boldsymbol{\Sigma}}_\beta^{-1}\tilde{\boldsymbol{\mu}}_\beta.$$

Therefore, the full conditional distribution of $\boldsymbol{\beta}$ is given by:

$$p(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\Gamma},\boldsymbol{\Sigma}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}}_\beta)^T\tilde{\boldsymbol{\Sigma}}_\beta^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}}_\beta)\right\},$$

which corresponds to a multivariate normal distribution with parameters specified by:

$$\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\Gamma},\boldsymbol{\Sigma} \sim \mathcal{N}_p(\tilde{\boldsymbol{\mu}}_\beta,\tilde{\boldsymbol{\Sigma}}_\beta), \tag{2.2.6}$$

where $\mathbf{W} = (W_1, W_2, \ldots, W_n)^T$, W_i is defined in Equation (2.2.4), $\tilde{\boldsymbol{\mu}}_\beta$ and $\tilde{\boldsymbol{\Sigma}}_\beta$ are defined in Equation (2.2.5). Since both the posterior mean $\tilde{\boldsymbol{\mu}}_\beta$ and covariance $\tilde{\boldsymbol{\Sigma}}_\beta$ can be derived in closed form, sampling $\boldsymbol{\beta}$ is computationally efficient within each iteration of the Gibbs sampler. This also serves as a natural closing point for the section.

2.2.4 Full Conditional Distribution for Γ

To derive the full conditional distribution for $\boldsymbol{\Gamma}$, we extract all terms from (2.2.3) that involve $\boldsymbol{\Gamma}$:

$$\begin{aligned} p(\boldsymbol{\Gamma}|\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\Sigma}) \propto \exp&\left\{-\frac{1}{2}\text{tr}\left[(\mathbf{X}-\mathbf{Z}\boldsymbol{\Gamma})\boldsymbol{\Sigma}_u^{-1}(\mathbf{X}-\mathbf{Z}\boldsymbol{\Gamma})^T\right]\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma^2_{\varepsilon|u}}\sum_{i=1}^{n}(Y_i-\mathbf{X}_i\boldsymbol{\beta}-\boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}(\mathbf{X}_i-\mathbf{Z}_i\boldsymbol{\Gamma})^T)^2\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma^2_\Gamma}\text{tr}[(\boldsymbol{\Gamma}-\boldsymbol{\mu}_\Gamma)^T(\boldsymbol{\Gamma}-\boldsymbol{\mu}_\Gamma)]\right\}. \end{aligned}$$

This expression consists of three main components: the first term relates to the measurement equation error structure, the second term captures the conditional error variance, and the third term represents the prior distribution for $\boldsymbol{\Gamma}$.

We begin our detailed analysis with the first exponential term by expanding the trace expression. The trace of the quadratic form can be decomposed as:

$$\text{tr}[(\mathbf{X}-\mathbf{Z}\boldsymbol{\Gamma})\boldsymbol{\Sigma}_u^{-1}(\mathbf{X}-\mathbf{Z}\boldsymbol{\Gamma})^T] = \text{tr}[\mathbf{X}\boldsymbol{\Sigma}_u^{-1}\mathbf{X}^T] - 2\text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{X}\boldsymbol{\Sigma}_u^{-1}] + \text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_u^{-1}].$$

Since $\text{tr}[\mathbf{X}\boldsymbol{\Sigma}_u^{-1}\mathbf{X}^T]$ doesn't depend on $\boldsymbol{\Gamma}$, we focus on the terms with $\boldsymbol{\Gamma}$:

$$-\frac{1}{2}\left[-2\text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{X}\boldsymbol{\Sigma}_u^{-1}] + \text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_u^{-1}]\right] = \text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{X}\boldsymbol{\Sigma}_u^{-1}] - \frac{1}{2}\text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_u^{-1}].$$

To simplify notation, the term $\text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{X}\boldsymbol{\Sigma}_u^{-1}]$ can be written as $\text{tr}[\boldsymbol{\Gamma}^T\mathbf{B}_1]$ where:

$$\mathbf{B}_1 = \mathbf{Z}^T\mathbf{X}\boldsymbol{\Sigma}_u^{-1} \tag{2.2.7}$$

.

For the quadratic term $\text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_u^{-1}]$, we can use vectorization properties to rewrite it as:

$$\text{tr}[\boldsymbol{\Gamma}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_u^{-1}] = \text{vec}(\boldsymbol{\Gamma})^T(\boldsymbol{\Sigma}_u^{-1}\otimes\mathbf{Z}^T\mathbf{Z})\text{vec}(\boldsymbol{\Gamma}).$$

This vectorization approach proves essential for combining all quadratic terms later in the derivation.

Next, we turn our attention to the second exponential term, which requires more careful treatment due to its complex structure. We begin by introducing convenient notation: let's define:

$$\begin{aligned}\varepsilon_i &= Y_i - \mathbf{X}_i\boldsymbol{\beta},\\ \mathbf{C} &= \boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}.\end{aligned}$$

This allows us to rewrite the second exponential term as:

$$\exp\left\{-\frac{1}{2\sigma_{\varepsilon|u}^2}\sum_{i=1}^{n}(\varepsilon_i - \mathbf{C}(\mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma})^T)^2\right\}.$$

To extract the $\boldsymbol{\Gamma}$-dependent terms, we expand the squared expression inside the summation:

$$\begin{aligned}(\varepsilon_i - \mathbf{C}(\mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma})^T)^2 &= \varepsilon_i^2 - 2\varepsilon_i\cdot\mathbf{C}(\mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma})^T + (\mathbf{C}(\mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma})^T)^2\\ &= \varepsilon_i^2 - 2\varepsilon_i\cdot\mathbf{C}\mathbf{X}_i^T + 2\varepsilon_i\cdot\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T + (\mathbf{C}\mathbf{X}_i^T)^2\\ &\quad - 2\mathbf{C}\mathbf{X}_i^T\cdot\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T + (\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T)^2.\end{aligned}$$

Among these terms, ε_i^2, $-2\varepsilon_i\cdot\mathbf{C}\mathbf{X}_i^T$, and $(\mathbf{C}\mathbf{X}_i^T)^2$ do not involve $\boldsymbol{\Gamma}$ and can be treated as constants. We focus on the terms with $\boldsymbol{\Gamma}$:

$$2\varepsilon_i\cdot\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T - 2\mathbf{C}\mathbf{X}_i^T\cdot\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T + (\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T)^2.$$

The linear terms can be combined and simplified as follows:

$$\begin{aligned}2\varepsilon_i\cdot\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T - 2\mathbf{C}\mathbf{X}_i^T\cdot\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T &= 2[\varepsilon_i - \mathbf{C}\mathbf{X}_i^T]\cdot\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T\\ &= 2[\varepsilon_i - \mathbf{C}\mathbf{X}_i^T]\cdot\mathbf{C}\boldsymbol{\Gamma}^T\mathbf{Z}_i^T.\end{aligned}$$

When we sum this expression over all observations and apply the trace operation, we obtain:

$$\sum_{i=1}^{n}2[\varepsilon_i - \mathbf{C}\mathbf{X}_i^T]\cdot\mathbf{C}\boldsymbol{\Gamma}^T\mathbf{Z}_i^T = 2\mathrm{tr}\left[\boldsymbol{\Gamma}^T\mathbf{Z}^T[\boldsymbol{\varepsilon} - \mathbf{X}\mathbf{C}^T]\mathbf{C}\right].$$

This can be written as $2\mathrm{tr}[\boldsymbol{\Gamma}^T\mathbf{B}_2]$ where:

$$\mathbf{B}_2 = \frac{1}{\sigma_{\varepsilon|u}^2}\mathbf{Z}^T[\boldsymbol{\varepsilon} - \mathbf{X}\mathbf{C}^T]\mathbf{C}. \tag{2.2.8}$$

For the quadratic term $(\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T)^2$, we require careful manipulation. We can rewrite this as:

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{C}(\mathbf{Z}_i\boldsymbol{\Gamma})^T)^2 &= \sum_{i=1}^{n}(\mathbf{C}\boldsymbol{\Gamma}^T\mathbf{Z}_i^T)^2\\ &= \sum_{i=1}^{n}\mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{C}^T\mathbf{C}\boldsymbol{\Gamma}^T\mathbf{Z}_i^T\\ &= \sum_{i=1}^{n}\mathrm{tr}[\mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{C}^T\mathbf{C}\boldsymbol{\Gamma}^T\mathbf{Z}_i^T]\\ &= \sum_{i=1}^{n}\mathrm{tr}[\boldsymbol{\Gamma}\mathbf{C}^T\mathbf{C}\boldsymbol{\Gamma}^T\mathbf{Z}_i^T\mathbf{Z}_i].\end{aligned}$$

Let's define:

$$\mathbf{Q}_i = \mathbf{Z}_i^T \mathbf{Z}_i,$$
$$\mathbf{R} = \mathbf{C}^T \mathbf{C}.$$

Both of these are symmetric matrices. Using the cyclic property of trace, we have:

$$\text{tr}[\boldsymbol{\Gamma}\mathbf{R}\boldsymbol{\Gamma}^T\mathbf{Q}_i] = \text{tr}[\boldsymbol{\Gamma}^T\mathbf{Q}_i\boldsymbol{\Gamma}\mathbf{R}].$$

Applying the vectorization property for traces of matrix products, where $\text{tr}[\mathbf{ABCD}] = \text{vec}(\mathbf{A}^T)^T(\mathbf{D}^T \otimes \mathbf{B})\text{vec}(\mathbf{C})$, we obtain:

$$\text{tr}[\boldsymbol{\Gamma}^T\mathbf{Q}_i\boldsymbol{\Gamma}\mathbf{R}] = \text{vec}(\boldsymbol{\Gamma})^T(\mathbf{R}^T \otimes \mathbf{Q}_i)\text{vec}(\boldsymbol{\Gamma}).$$

Since $\mathbf{R} = \mathbf{C}^T\mathbf{C}$ is a symmetric matrix, $\mathbf{R}^T = \mathbf{R}$. Therefore:

$$\sum_{i=1}^{n} \text{tr}[\boldsymbol{\Gamma}\mathbf{R}\boldsymbol{\Gamma}^T\mathbf{Q}_i] = \sum_{i=1}^{n} \text{vec}(\boldsymbol{\Gamma})^T(\mathbf{R} \otimes \mathbf{Q}_i)\text{vec}(\boldsymbol{\Gamma}).$$

Moving on to analyze the prior term, we consider the third exponential term which represents the prior distribution for $\boldsymbol{\Gamma}$:

$$\exp\left\{-\frac{1}{2\sigma_\Gamma^2}\text{tr}[(\boldsymbol{\Gamma} - \boldsymbol{\mu}_\Gamma)^T(\boldsymbol{\Gamma} - \boldsymbol{\mu}_\Gamma)]\right\}.$$

Expanding the quadratic form in the exponent:

$$\text{tr}[(\boldsymbol{\Gamma} - \boldsymbol{\mu}_\Gamma)^T(\boldsymbol{\Gamma} - \boldsymbol{\mu}_\Gamma)] = \text{tr}[\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} - 2\boldsymbol{\Gamma}^T\boldsymbol{\mu}_\Gamma + \boldsymbol{\mu}_\Gamma^T\boldsymbol{\mu}_\Gamma].$$

Since $\text{tr}[\boldsymbol{\mu}_\Gamma^T\boldsymbol{\mu}_\Gamma]$ does not depend on $\boldsymbol{\Gamma}$, we focus on:

$$\text{tr}[\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}] - 2\text{tr}[\boldsymbol{\Gamma}^T\boldsymbol{\mu}_\Gamma].$$

The linear term $2\text{tr}[\boldsymbol{\Gamma}^T\boldsymbol{\mu}_\Gamma]$ can be written as $2\text{tr}[\boldsymbol{\Gamma}^T\mathbf{B}_3]$ where:

$$\mathbf{B}_3 = \frac{1}{\sigma_\Gamma^2}\boldsymbol{\mu}_\Gamma. \tag{2.2.9}$$

For the quadratic term, $\text{tr}[\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}]$ can be written using vectorization as:

$$\text{tr}[\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}] = \text{vec}(\boldsymbol{\Gamma})^T(\mathbf{I}_p \otimes \mathbf{I}_q)\text{vec}(\boldsymbol{\Gamma}).$$

where $\mathbf{I}_p$ and $\mathbf{I}_q$ are identity matrices of appropriate dimensions.

Having analyzed each component separately, we now combine all the $\boldsymbol{\Gamma}$-dependent terms. The complete expression becomes:

$$\text{tr}[\boldsymbol{\Gamma}^T(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3)] - \frac{1}{2}\text{vec}(\boldsymbol{\Gamma})^T(\boldsymbol{\Sigma}_u^{-1} \otimes \mathbf{Z}^T\mathbf{Z})\text{vec}(\boldsymbol{\Gamma})$$
$$- \frac{1}{2\sigma_{\varepsilon|u}^2}\sum_{i=1}^{n} \text{vec}(\boldsymbol{\Gamma})^T(\mathbf{R} \otimes \mathbf{Q}_i)\text{vec}(\boldsymbol{\Gamma}) - \frac{1}{2\sigma_\Gamma^2}\text{vec}(\boldsymbol{\Gamma})^T(\mathbf{I}_p \otimes \mathbf{I}_q)\text{vec}(\boldsymbol{\Gamma}).$$

We can simplify the quadratic coefficient by noting that $\sum_{i=1}^n \mathbf{Q}_i = \sum_{i=1}^n \mathbf{Z}_i^T\mathbf{Z}_i = \mathbf{Z}^T\mathbf{Z} = \mathbf{Q}$. Using the property that $\text{tr}[\mathbf{A}^T\mathbf{B}] = \text{vec}(\mathbf{A})^T\text{vec}(\mathbf{B})$:

$$\text{tr}[\boldsymbol{\Gamma}^T(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3)] = \text{vec}(\boldsymbol{\Gamma})^T\text{vec}(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3).$$

Let's define the precision matrix:

$$\begin{aligned}\boldsymbol{\Lambda} &= (\boldsymbol{\Sigma}_u^{-1} \otimes \mathbf{Q}) + \frac{1}{\sigma^2_{\varepsilon|u}} \sum_{i=1}^{n} (\mathbf{R} \otimes \mathbf{Q}_i) + \frac{1}{\sigma^2_{\Gamma}} (\mathbf{I}_p \otimes \mathbf{I}_q) \\ &= (\boldsymbol{\Sigma}_u^{-1} \otimes \mathbf{Q}) + \frac{1}{\sigma^2_{\varepsilon|u}} (\mathbf{R} \otimes \mathbf{Q}) + \frac{1}{\sigma^2_{\Gamma}} (\mathbf{I}_p \otimes \mathbf{I}_q),\end{aligned} \tag{2.2.10}$$

where $\mathbf{Q} = \mathbf{Z}^T\mathbf{Z}$.
The complete exponent now takes the form:

$$\text{vec}(\boldsymbol{\Gamma})^T \text{vec}(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3) - \frac{1}{2} \text{vec}(\boldsymbol{\Gamma})^T \boldsymbol{\Lambda} \text{vec}(\boldsymbol{\Gamma}).$$

To complete the square, we rewrite this expression as:

$$\begin{aligned}&\text{vec}(\boldsymbol{\Gamma})^T \text{vec}(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3) - \frac{1}{2} \text{vec}(\boldsymbol{\Gamma})^T \boldsymbol{\Lambda} \text{vec}(\boldsymbol{\Gamma}) \\ &= -\frac{1}{2} [\text{vec}(\boldsymbol{\Gamma})^T \boldsymbol{\Lambda} \text{vec}(\boldsymbol{\Gamma}) - 2\text{vec}(\boldsymbol{\Gamma})^T \text{vec}(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3)] \\ &= -\frac{1}{2} [\text{vec}(\boldsymbol{\Gamma})^T \boldsymbol{\Lambda} \text{vec}(\boldsymbol{\Gamma}) - 2\text{vec}(\boldsymbol{\Gamma})^T \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1} \text{vec}(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3)] \\ &= -\frac{1}{2} [(\text{vec}(\boldsymbol{\Gamma}) - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\text{vec}(\boldsymbol{\Gamma}) - \boldsymbol{\mu})] + \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{\mu},\end{aligned}$$

where

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1} \text{vec}(\mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3). \tag{2.2.11}$$

The completed square form reveals that the full conditional distribution for $\text{vec}(\boldsymbol{\Gamma})$ follows a multivariate normal distribution. Therefore, the full conditional distribution for $\text{vec}(\boldsymbol{\Gamma})$ is:

$$\text{vec}(\boldsymbol{\Gamma}) | \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_{qp}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \tag{2.2.12}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are defined by equations (2.2.11) and (2.2.10), with $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{B}_3$ given by equations (2.2.7), (2.2.8), and (2.2.9). This derivation demonstrates how the complex interaction between the measurement equation, conditional error structure, and prior information combines to yield a tractable posterior distribution suitable for Gibbs sampling implementations.

2.2.5 Full Conditional Distribution for Σ

To derive the full conditional distribution for $\boldsymbol{\Sigma}$, we extract all terms from (2.2.3) that involve $\boldsymbol{\Sigma}$:

$$\begin{aligned}p(\boldsymbol{\Sigma} | \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}) \propto\ & |\boldsymbol{\Sigma}_u|^{-n/2} (\sigma^2_{\varepsilon|u})^{-n/2} \\ & \times \exp\left\{ -\frac{1}{2} \left[\text{tr}\left[(\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma}) \boldsymbol{\Sigma}_u^{-1} (\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma})^T \right] \right.\right. \\ & \left.\left. + \frac{1}{\sigma^2_{\varepsilon|u}} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} (\mathbf{X}_i - \mathbf{Z}_i\boldsymbol{\Gamma})^T)^2 \right] \right\} \\ & \times |\boldsymbol{\Psi}_0|^{\nu_0/2} |\boldsymbol{\Sigma}|^{-(\nu_0+p+2)/2} \exp\left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Psi}_0 \boldsymbol{\Sigma}^{-1}) \right\}.\end{aligned} \tag{2.2.13}$$

First, we recall that $\boldsymbol{\Sigma}$ has a block structure (2.1.1), using properties of block matrices, the determinant can be written as:

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_u| \cdot |\sigma_\varepsilon^2 - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon}| = |\boldsymbol{\Sigma}_u| \cdot \sigma_{\varepsilon|u}^2,$$

where $\sigma_{\varepsilon|u}^2 = \sigma_\varepsilon^2 - \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon}$.

The inverse of the block matrix is shown in (2.2.1). Combining the determinant terms from (2.2.13), we have:

$$|\boldsymbol{\Sigma}_u|^{-n/2} \cdot (\sigma_{\varepsilon|u}^2)^{-n/2} \cdot |\boldsymbol{\Sigma}|^{-(\nu_0+p+2)/2}.$$

Using the relation $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_u| \cdot \sigma_{\varepsilon|u}^2$, we can rewrite this as:

$$|\boldsymbol{\Sigma}|^{-n/2} \cdot |\boldsymbol{\Sigma}|^{-(\nu_0+p+2)/2} = |\boldsymbol{\Sigma}|^{\cdot}$$

Next, we define the residuals:

$$\begin{aligned} \varepsilon_i &= Y_i - \mathbf{X}_i \boldsymbol{\beta} \quad \text{(second-stage residual)}, \\ \mathbf{U}_i &= (\mathbf{X}_i - \mathbf{Z}_i \boldsymbol{\Gamma})^T \quad \text{(first-stage residual as a column vector)}, \\ \mathbf{R}_i &= \begin{pmatrix} \mathbf{U}_i \\ \varepsilon_i \end{pmatrix} \quad \text{(combined residual vector)}. \end{aligned} \tag{2.2.14}$$

We can compute the matrix product $\mathbf{R}_i \mathbf{R}_i^T$:

$$\mathbf{R}_i \mathbf{R}_i^T = \begin{pmatrix} \mathbf{U}_i \\ \varepsilon_i \end{pmatrix} \begin{pmatrix} \mathbf{U}_i^T & \varepsilon_i \end{pmatrix} = \begin{pmatrix} \mathbf{U}_i \mathbf{U}_i^T & \mathbf{U}_i \varepsilon_i \\ \varepsilon_i \mathbf{U}_i^T & \varepsilon_i^2 \end{pmatrix}.$$

Now, we need to compute $\operatorname{tr}(\mathbf{R}_i \mathbf{R}_i^T \boldsymbol{\Sigma}^{-1})$. Using the block structure of $\boldsymbol{\Sigma}^{-1}$ from (2.2.1) and the property that for block matrices:

$$\operatorname{tr}\left(\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} \end{pmatrix} \right) = \operatorname{tr}(\mathbf{AE} + \mathbf{BG}) + \operatorname{tr}(\mathbf{CF} + \mathbf{DH}),$$

We have:

$$\begin{aligned} \operatorname{tr}(\mathbf{R}_i \mathbf{R}_i^T \boldsymbol{\Sigma}^{-1}) = \operatorname{tr} &\left(\mathbf{U}_i \mathbf{U}_i^T \left(\boldsymbol{\Sigma}_u^{-1} + \frac{1}{\sigma_{\varepsilon|u}^2} \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon} \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \right) \right) \\ &+ \operatorname{tr} \left(\mathbf{U}_i \varepsilon_i \left(-\frac{1}{\sigma_{\varepsilon|u}^2} \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \right) \right) \\ &+ \operatorname{tr} \left(\varepsilon_i \mathbf{U}_i^T \left(-\frac{1}{\sigma_{\varepsilon|u}^2} \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon} \right) \right) \\ &+ \operatorname{tr} \left(\varepsilon_i^2 \frac{1}{\sigma_{\varepsilon|u}^2} \right). \end{aligned}$$

For the second and third terms, since $\mathbf{U}_i$ is a $p \times 1$ column vector, ε_i is a scalar, and the matrix multiplications result in scalars for these trace operations:

$$\begin{aligned} \operatorname{tr} \left(\mathbf{U}_i \varepsilon_i \left(-\frac{1}{\sigma_{\varepsilon|u}^2} \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \right) \right) &= -\frac{1}{\sigma_{\varepsilon|u}^2} \varepsilon_i \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \mathbf{U}_i, \\ \operatorname{tr} \left(\varepsilon_i \mathbf{U}_i^T \left(-\frac{1}{\sigma_{\varepsilon|u}^2} \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\Sigma}_{u\varepsilon} \right) \right) &= -\frac{1}{\sigma_{\varepsilon|u}^2} \varepsilon_i \boldsymbol{\Sigma}_{u\varepsilon}^T \boldsymbol{\Sigma}_u^{-1} \mathbf{U}_i. \end{aligned}$$

The fourth term is:

$$\mathrm{tr}\left(\varepsilon_i^2 \frac{1}{\sigma^2_{\varepsilon|u}}\right) = \frac{\varepsilon_i^2}{\sigma^2_{\varepsilon|u}} = \frac{(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2}{\sigma^2_{\varepsilon|u}}.$$

Combining all terms:

$$\begin{aligned}\mathrm{tr}(\mathbf{R}_i\mathbf{R}_i^T\boldsymbol{\Sigma}^{-1}) = \mathrm{tr}\left(\mathbf{U}_i\mathbf{U}_i^T\boldsymbol{\Sigma}_u^{-1}\right) \\ + \frac{1}{\sigma^2_{\varepsilon|u}}\Big[(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 - 2(Y_i - \mathbf{X}_i\boldsymbol{\beta})\boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}\mathbf{U}_i \\ + \mathbf{U}_i^T\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_{u\varepsilon}\boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}\mathbf{U}_i\Big]\end{aligned}$$

Recognizing the perfect square pattern in the bracketed term:

$$(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 - 2(Y_i - \mathbf{X}_i\boldsymbol{\beta})\boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}\mathbf{U}_i + \mathbf{U}_i^T\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_{u\varepsilon}\boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}\mathbf{U}_i = (Y_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}\mathbf{U}_i)^2,$$

therefore we can get:

$$\mathrm{tr}(\mathbf{R}_i\mathbf{R}_i^T\boldsymbol{\Sigma}^{-1}) = \mathrm{tr}\left(\mathbf{U}_i\mathbf{U}_i^T\boldsymbol{\Sigma}_u^{-1}\right) + \frac{(Y_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}\mathbf{U}_i)^2}{\sigma^2_{\varepsilon|u}}.$$

Summing over all i and using the linearity of the trace:

$$\mathrm{tr}\left[\sum_{i=1}^n \mathbf{R}_i\mathbf{R}_i^T\boldsymbol{\Sigma}^{-1}\right] = \mathrm{tr}\left[\sum_{i=1}^n \mathbf{U}_i\mathbf{U}_i^T\boldsymbol{\Sigma}_u^{-1}\right] + \sum_{i=1}^n \frac{(Y_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Sigma}_{u\varepsilon}^T\boldsymbol{\Sigma}_u^{-1}\mathbf{U}_i)^2}{\sigma^2_{\varepsilon|u}}.$$

This proves that our original two expressions from the likelihood can be combined into a single matrix trace form.

Combining all terms, along with the prior term $\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1})\right\}$, we get:

$$p(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\Gamma}) \propto |\boldsymbol{\Sigma}|^{-(n+\nu_0+p+2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left[\sum_{i=1}^n \mathbf{R}_i\mathbf{R}_i^T + \boldsymbol{\Psi}_0\right]\boldsymbol{\Sigma}^{-1}\right)\right\}.$$

This is precisely the form of the density function of an Inverse Wishart distribution:

$$\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\Gamma} \sim \text{Inverse-Wishart}_{p+1}\left(n + \nu_0, \sum_{i=1}^n \mathbf{R}_i\mathbf{R}_i^T + \boldsymbol{\Psi}_0\right). \tag{2.2.15}$$

where $n + \nu_0$ is the degrees of freedom, $\sum_{i=1}^n \mathbf{R}_i\mathbf{R}_i^T + \boldsymbol{\Psi}_0$ is the scale matrix, and $\mathbf{R}_i$ is defined in (2.2.14).

2.2.6 Initial Simulation Results

To evaluate the performance of Full Bayesian, we conducted a simulation study using synthetic data generated from a manually specified parameter configuration. The data were simulated under a multivariate IV setting with $p = 2$ (number of endogenous covariates), $q = 3$ (number of instruments), and sample size $n = 200$. The true parameters were set as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\Gamma} = \begin{bmatrix} 1.0 & 0.5 \\ 0.3 & -0.2 \\ 0.0 & 0.8 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.5 & -0.4 \\ 0.5 & 1.0 & -0.4 \\ -0.4 & -0.4 & 1.0 \end{bmatrix}.$$

Here, $\boldsymbol{\Gamma}$ represents the first-stage coefficients and was fixed to reflect a heterogeneous instrument relevance structure. The covariance matrix $\boldsymbol{\Sigma}$ was specified as a block structure (2.1.1) incorporating both the covariance of the first-stage errors and their correlation with the structural error term.

Posterior inference under Full Bayesian was carried out using a Gibbs sampler with 5000 iterations, including a burn-in period of 1000 iterations. The posterior mean estimates are summarized in Table 2.2.1. While the posterior means of $\boldsymbol{\beta}$ were relatively accurate—deviating by only 0.04 and 0.57 from the true values—the estimates of $\boldsymbol{\Sigma}$ showed noticeable overestimation. For instance, the off-diagonal entry Σ_{13} was estimated as -0.8516 compared to its true value of -0.4, and Σ_{33} was overestimated as 2.3433 versus the true 1.0. The posterior mean of $\boldsymbol{\Gamma}$ deviated more substantially: the largest difference between the estimated and true values was as high as 0.6, and the overall deviation across all entries was also substantial. This indicates that, under the full Bayesian model, the recovery of the first-stage coefficients $\boldsymbol{\Gamma}$ is notably less accurate.

Table 2.2.1: True values vs posterior means of $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ under Full Bayesian

Parameter	True Value	Posterior Mean
$\boldsymbol{\beta}$	$\begin{bmatrix} 1.00 \\ 1.00 \end{bmatrix}$	$\begin{bmatrix} 0.9629 \\ 1.5687 \end{bmatrix}$
$\boldsymbol{\Gamma}$	$\begin{bmatrix} 1.0 & 0.5 \\ 0.3 & -0.2 \\ 0.0 & 0.8 \end{bmatrix}$	$\begin{bmatrix} 0.7204 & 0.0258 \\ 0.3905 & -0.0517 \\ -0.3170 & 0.2023 \end{bmatrix}$
$\boldsymbol{\Sigma}$	$\begin{bmatrix} 1.0 & 0.5 & -0.4 \\ 0.5 & 1.0 & -0.4 \\ -0.4 & -0.4 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 1.0979 & 0.9194 & -0.8516 \\ 0.9194 & 1.8031 & -1.5564 \\ -0.8516 & -1.5564 & 2.3433 \end{bmatrix}$

To further assess posterior behavior, we examined the trace plots of selected parameters, shown in Figure 2.2.1. The sampling chains for β_1, β_2, and Σ_{33} exhibited reasonable stationarity and variability. However, the trace plots of β_2 and Σ_{33} indicated poor convergence.

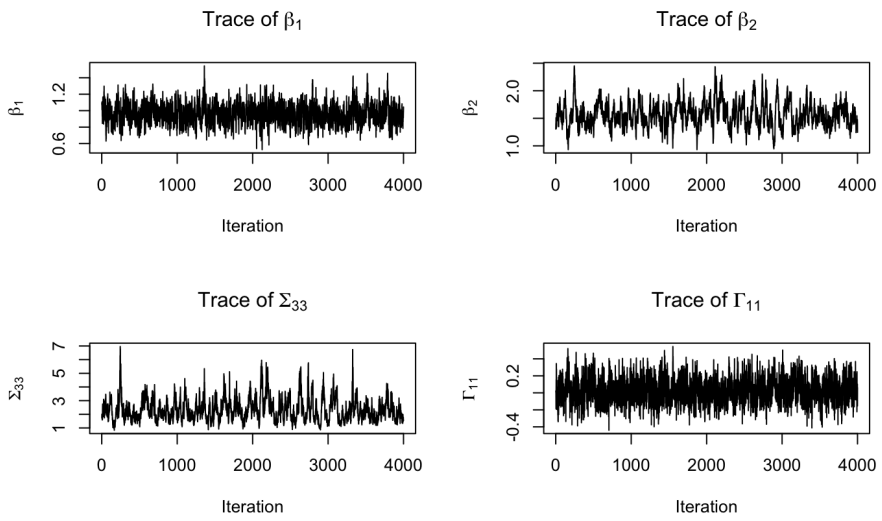

Figure 2.2.1: Trace plots of selected parameters under Full Bayesian. While chains for $\boldsymbol{\beta}$ and Σ_{33} show adequate mixing, the narrow fluctuation band for Γ_{11} indicates potential under-mixing.

These results suggest that the joint sampling of $\mathbf{\Gamma}$ in the full Bayesian specification may induce strong dependencies with the structural parameters, especially in finite samples. To mitigate this issue, the simplified Bayesian approach (Simplified Bayesian), discussed in Section 2.3, estimates $\mathbf{\Gamma}$ from the first-stage equation independently. This modularization reduces posterior entanglement and improves sampling efficiency and robustness in sparse or weakly identified designs.

2.3 Simplified Bayesian Approach

2.3.1 Motivation from Simulation Findings

For full Bayesian approach, our preliminary simulation studies revealed several practical challenges in its implementation. Most notably, the posterior estimates for the first-stage coefficient matrix $\mathbf{\Gamma}$ deviated considerably from the true values, indicating that the full Bayesian model had difficulty capturing $\mathbf{\Gamma}$ accurately.

Furthermore, the precision matrix $\mathbf{\Lambda} \in \mathbb{R}^{pq \times pq}$ defined in equation (2.2.10) requires matrix inversion at each MCMC iteration. As $\mathbf{\Lambda}$ is a dense matrix of dimension $pq \times pq$, this cost can become substantial when the number of instruments q or endogenous variables p is large, adding considerable computational burden to the full Bayesian approach.

These computational challenges motivated us to develop a simplified approach that maintains the essential structure for correcting endogeneity and convergence properties.

2.3.2 Modified Posterior Framework

The full conditional distribution for $\mathbf{\Gamma}$ in full Bayesian approach, as derived in equation (2.2.12), depends on all observed variables $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$, as well as all other

model parameters. Simplified Bayesian approach addresses the computational challenges of full Bayesian approach by modifying how we estimate the first-stage coefficients $\mathbf{\Gamma}$. The key insight is that we can separate the estimation of $\mathbf{\Gamma}$ from the full joint posterior by conditioning only on the first-stage relationship between the endogenous variables $\mathbf{X}$ and instruments $\mathbf{Z}$.

In simplified Bayesian approach, we estimate $\mathbf{\Gamma}$ using only the first-stage equation:

$$\mathbf{X} = \mathbf{Z\Gamma} + \mathbf{U}$$

without directly incorporating information from the second-stage equation. This approach is conceptually similar to the first stage of two-stage least squares (2SLS) in the frequentist framework, but maintains the Bayesian structure for uncertainty quantification.

For the second-stage parameters $\boldsymbol{\beta}$ and the covariance matrix $\mathbf{\Sigma}$, Simplified Bayesian retains the same full conditional distributions as Full Bayesian, ensuring that the correction for endogeneity is preserved through the joint modeling of the error terms.

2.3.3 Modified Full Conditional Distribution for Γ

In Simplified Bayesian, the full conditional distribution for $\mathbf{\Gamma}$ is derived using only the first-stage likelihood and the prior distribution:

$$p(\mathbf{\Gamma}|\mathbf{X}, \mathbf{Z}, \mathbf{\Sigma}_u) \propto p(\mathbf{X}|\mathbf{Z}, \mathbf{\Gamma}, \mathbf{\Sigma}_u) \times p(\mathbf{\Gamma}).$$

The first-stage likelihood is:

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{\Gamma}, \mathbf{\Sigma}_u) \propto |\mathbf{\Sigma}_u|^{-n/2} \exp\left\{-\frac{1}{2}\text{tr}\left[(\mathbf{X} - \mathbf{Z\Gamma})\mathbf{\Sigma}_u^{-1}(\mathbf{X} - \mathbf{Z\Gamma})^T\right]\right\},$$

combined with the prior $\mathbf{\Gamma} \sim \mathcal{N}(\boldsymbol{\mu}_\Gamma, \sigma_\Gamma^2 \mathbf{I}_{q\times p})$, we obtain:

$$\begin{aligned} p(\mathbf{\Gamma}|\mathbf{X}, \mathbf{Z}, \mathbf{\Sigma}_u) \propto &\exp\left\{-\frac{1}{2}\text{tr}\left[(\mathbf{X} - \mathbf{Z\Gamma})\mathbf{\Sigma}_u^{-1}(\mathbf{X} - \mathbf{Z\Gamma})^T\right]\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma_\Gamma^2}\text{tr}\left[(\mathbf{\Gamma} - \boldsymbol{\mu}_\Gamma)^T(\mathbf{\Gamma} - \boldsymbol{\mu}_\Gamma)\right]\right\}. \end{aligned}$$

Expanding the trace terms and collecting those involving $\mathbf{\Gamma}$:

$$p(\mathbf{\Gamma}|\mathbf{X}, \mathbf{Z}, \mathbf{\Sigma}_u) \propto \exp\left\{-\frac{1}{2}\text{tr}\left[\mathbf{\Gamma}^T\mathbf{Z}^T\mathbf{Z\Gamma}\mathbf{\Sigma}_u^{-1} - 2\mathbf{X}^T\mathbf{Z\Gamma}\mathbf{\Sigma}_u^{-1} + \frac{1}{\sigma_\Gamma^2}(\mathbf{\Gamma} - \boldsymbol{\mu}_\Gamma)^T(\mathbf{\Gamma} - \boldsymbol{\mu}_\Gamma)\right]\right\}.$$

Using properties of the matrix-normal distribution, this can be shown to be:

$$\mathbf{\Gamma}|\mathbf{X}, \mathbf{Z}, \mathbf{\Sigma}_u \sim \text{Matrix-Normal}_{q\times p}(\tilde{\boldsymbol{\mu}}_\Gamma, \tilde{\mathbf{V}}_\Gamma, \mathbf{\Sigma}_u),$$

where:

$$\begin{aligned} \tilde{\mathbf{V}}_\Gamma &= \left(\mathbf{Z}^T\mathbf{Z} + \frac{1}{\sigma_\Gamma^2}\mathbf{I}_q\right)^{-1}, \\ \tilde{\boldsymbol{\mu}}_\Gamma &= \tilde{\mathbf{V}}_\Gamma\left(\mathbf{Z}^T\mathbf{X} + \frac{1}{\sigma_\Gamma^2}\boldsymbol{\mu}_\Gamma\right). \end{aligned}$$

Note that this full conditional distribution is significantly simpler than the corresponding distribution in full Bayesian approach, as it does not depend on $\mathbf{Y}$, $\boldsymbol{\beta}$, or the cross-covariance terms in $\mathbf{\Sigma}$.

2.3.4 Full Conditional Distribution for β and Σ

For the second-stage coefficient vector $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$, Simplified Bayesian uses the same full conditional distributions as derived for Full Bayesian. Specifically, the full conditional distribution for $\boldsymbol{\beta}$ follows the multivariate normal distribution given in equation (2.2.6), and the full conditional distribution for $\boldsymbol{\Sigma}$ follows the inverse Wishart distribution given in equation (2.2.15).

In full Bayesian approach, this retention of full conditional distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ is crucial because they capture the essential endogeneity correction mechanism. The correlation between the first-stage errors $\mathbf{U}$ and the second-stage errors $\boldsymbol{\varepsilon}$, as encoded in the covariance matrix $\boldsymbol{\Sigma}$, allows for appropriate adjustment of endogeneity when estimating $\boldsymbol{\beta}$.

The key difference in the simplified Bayesian approach is that the full conditional distribution of $\boldsymbol{\Gamma}$ is independent of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, relying solely on the first-stage information from $\mathbf{X}$ and $\mathbf{Z}$. In contrast, the updates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ still incorporate the full joint structure through the definitions of W_i in equation (2.2.4) and the residual vectors $\mathbf{R}_i$ in equation (2.2.14). This design maintains endogeneity correction while simplifying the computation of first-stage parameters.

2.3.5 Initial Simulation Results

To evaluate the practical benefits of Simplified Bayesian, we conducted a simulation using the same data-generating process as described in Section 2.2.6.

Posterior inference was performed using a Gibbs sampler with 4000 iterations. Table 2.3.1 reports the true values alongside the posterior means obtained under simplified Bayesian approach. The estimates for $\boldsymbol{\Gamma}$ align closely with the true values, both in direction and magnitude, reflecting a clear improvement over the estimates obtained under the full Bayesian approach specification. In addition, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ remain close to the ground truth, confirming that the modularization in simplified Bayesian approach does not compromise the quality of inference for the remaining model components.

Table 2.3.1: True values vs posterior means of $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ under Simplified Bayesian

Parameter	True Value	Posterior Mean
$\boldsymbol{\beta}$	$\begin{bmatrix} 1.00 \\ 1.00 \end{bmatrix}$	$\begin{bmatrix} 0.9127 \\ 1.0821 \end{bmatrix}$
$\boldsymbol{\Gamma}$	$\begin{bmatrix} 1.0 & 0.5 \\ 0.3 & -0.2 \\ 0.0 & 0.8 \end{bmatrix}$	$\begin{bmatrix} 0.9746 & 0.4807 \\ 0.3328 & -0.1631 \\ 0.0111 & 0.8081 \end{bmatrix}$
$\boldsymbol{\Sigma}$	$\begin{bmatrix} 1.0 & 0.5 & -0.4 \\ 0.5 & 1.0 & -0.4 \\ -0.4 & -0.4 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 0.9318 & 0.6202 & -0.3208 \\ 0.6202 & 1.2536 & -0.5828 \\ -0.3208 & -0.5828 & 1.1137 \end{bmatrix}$

We further assess the convergence and mixing behavior of the MCMC sampler through trace plots presented in Figure 2.3.1. The sampling chains for β_1, β_2, Σ_{33}, and Γ_{11} all exhibit stable behavior, without drift or pathological patterns. Notably,

the trace plots of β_2 and Σ_{33} no longer exhibit poor convergence. Instead, both parameters display stable mixing behavior and symmetric fluctuations around their posterior means, indicating improved posterior exploration.

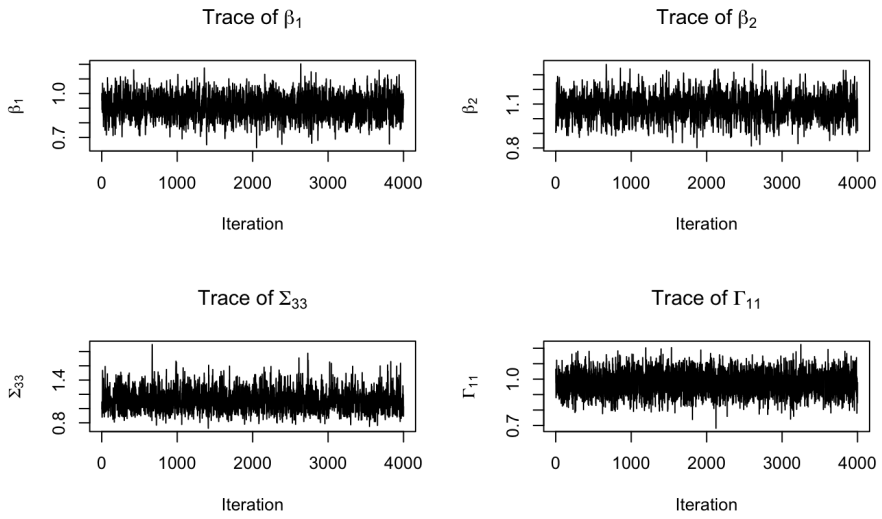

Figure 2.3.1: Trace plots of selected parameters under Simplified Bayesian. All chains show stable mixing and no signs of nonstationarity.

These results reinforce the advantage of simplifying the posterior sampling of $\mathbf{\Gamma}$ by conditioning only on the first-stage regression, $(\mathbf{X}, \mathbf{Z})$. By decoupling $\mathbf{\Gamma}$ from the full system and reducing posterior dependence, Simplified Bayesian achieves faster convergence and more accurate estimation in finite samples, particularly in settings where identification may be weak or partial.

2.4 Comprehensive Simulation Studies

2.4.1 Simulation Design

To evaluate the performance of our proposed Bayesian instrumental variables approaches, we conduct extensive simulation studies comparing three methods: the one-stage Bayesian regression (Naive), the full Bayesian approach (Full Bayesian), and our approach (Simplified Bayesian). The simulation design systematically examines various configurations of instrument and endogenous variable dimensions, alongside different sample sizes.

We generate data according to the following structural model (1.2.1). For each simulation replication, we set the true parameter $\boldsymbol{\beta}_{true} = (1, 1, \ldots, 1)^T$, which is a $p \times 1$ vector of ones, The true first-stage coefficient matrix $\mathbf{\Gamma}_{\text{true}}$ is re-sampled in each simulation from a uniform distribution Uniform$(0, 1)$ element-wise. This ensures that the simulated data spans a diverse set of instrument–endogeneity relationships.The covariance matrix $\mathbf{\Sigma}_u$ is set to have an AR(1) structure with $(\mathbf{\Sigma}_u)_{ij} = 0.5^{|i-j|}$, $\sigma_\varepsilon^2 = 1$, and the covariance vector $\mathbf{\Sigma}_{u\varepsilon} = (-0.4, -0.4, \ldots, -0.4)^T$. We examine three configurations representing different relationships between the number of instruments (q) and endogenous variables (p): $p > q$ $(p = 5, q = 3)$, $p = q$

$(p = 3, q = 3)$, and $p < q$ $(p = 2, q = 3)$. For each configuration, we consider sample sizes $n = 100$ and $n = 500$.

Each scenario involves 100 Monte Carlo replications. For each replication, we generate one dataset and apply all three methods to ensure fair comparison. We extract posterior means and 95% credible intervals for β and calculate performance metrics. For Bayesian methods, we use Gibbs sampling with 5,000 iterations and 1,000 burn-in iterations.

We evaluate performance using bias, mean squared error, and coverage probability:

$$\begin{aligned} \text{Bias}(\hat{\beta}_j) &= \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_{j,r} - \beta_{j,true}) \\ \text{MSE}(\hat{\beta}_j) &= \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_{j,r} - \beta_{j,true})^2 \\ \text{Coverage}(\beta_j) &= \frac{1}{R}\sum_{r=1}^{R}\mathbf{1}(\beta_{j,true} \in CI_{j,r}) \end{aligned} \tag{2.4.1}$$

where $R = 100$, $\hat{\beta}_{j,r}$ is the posterior mean in replication r, and $CI_{j,r}$ is the 95% credible interval.

2.4.2 Performance Across Sample Sizes and Different Dimensions

To assess how the three Bayesian IV methods behave under varying identification structures and sample sizes, we summarize the simulation outcomes in the tables below. For each configuration, we report the average bias, mean squared error (MSE), and coverage probability across all coefficients. The results are organized by sample size and categorized into over-identified, exactly-identified, and under-identified settings. This layout allows for direct comparison of the methods' performance across different empirical scenarios, highlighting both their strengths and potential limitations.

Table 2.4.1: Simulation results for $p = 2$, $q = 3$

n	Metric	Method	β_1	β_2
100	Bias	Naive	−0.1503	−0.1313
		Full Bayesian	0.1270	0.1972
		Simplified Bayesian	−0.0525	−0.0119
	MSE	Naive	0.0373	0.0334
		Full Bayesian	0.1299	0.1396
		Simplified Bayesian	0.0568	0.0443
	Coverage	Naive	0.67	0.65
		Full Bayesian	0.79	0.76
		Simplified Bayesian	0.94	0.96
500	Bias	Naive	−0.1470	−0.1243
		Full Bayesian	0.5448	0.6446
		Simplified Bayesian	−0.0247	0.0097
	MSE	Naive	0.0266	0.0221
		Full Bayesian	0.3996	0.5719
		Simplified Bayesian	0.0174	0.0184
	Coverage	Naive	0.17	0.27
		Full Bayesian	0.25	0.13
		Simplified Bayesian	0.95	0.93

Table 2.4.2: Simulation results for $p = 3$, $q = 3$

n	Metric	Method	β_1	β_2	β_3
100	Bias	Naive	−0.1509	−0.0626	−0.1158
		Full Bayesian	0.1334	0.1340	0.1819
		Simplified Bayesian	−0.0724	0.0396	−0.0424
	MSE	Naive	0.0386	0.0262	0.0300
		Full Bayesian	0.1138	0.1417	0.1302
		Simplified Bayesian	0.0507	0.0491	0.0452
	Coverage	Naive	0.56	0.78	0.73
		Full Bayesian	0.79	0.83	0.77
		Simplified Bayesian	0.98	0.92	0.97
500	Bias	Naive	−0.1333	−0.0710	−0.1214
		Full Bayesian	0.4276	0.2473	0.4467
		Simplified Bayesian	−0.0111	0.0026	−0.0069
	MSE	Naive	0.0251	0.0157	0.0225
		Full Bayesian	0.3397	0.3038	0.3500
		Simplified Bayesian	0.0229	0.0250	0.0187
	Coverage	Naive	0.27	0.54	0.30
		Full Bayesian	0.25	0.47	0.17
		Simplified Bayesian	0.90	0.91	0.93

Table 2.4.3: Simulation results for $p = 5$, $q = 3$

n	Metric	Method	β_1	β_2	β_3	β_4	β_5
100	Bias	Naive	−0.1239	−0.0623	−0.0573	−0.0839	−0.1075
		Full Bayesian	0.1294	0.0674	0.0687	−0.0086	0.1744
		Simplified Bayesian	−0.0513	−0.0158	−0.0339	−0.0311	−0.0300
	MSE	Naive	0.0312	0.0258	0.0276	0.0287	0.0279
		Full Bayesian	0.1004	0.1207	0.1436	0.1198	0.1110
		Simplified Bayesian	0.0270	0.0427	0.0422	0.0414	0.0291
	Coverage	Naive	0.68	0.84	0.77	0.77	0.72
		Full Bayesian	0.84	0.85	0.76	0.86	0.81
		Simplified Bayesian	1.00	0.98	0.99	0.98	0.99
500	Bias	Naive	−0.1284	−0.0503	−0.0569	−0.0887	−0.1040
		Full Bayesian	0.2404	0.1737	0.1487	0.0442	0.3367
		Simplified Bayesian	−0.0469	0.0014	−0.0154	−0.0503	−0.0182
	MSE	Naive	0.0266	0.0153	0.0184	0.0228	0.0192
		Full Bayesian	0.1611	0.1842	0.1890	0.1764	0.2148
		Simplified Bayesian	0.0227	0.0182	0.0200	0.0244	0.0175
	Coverage	Naive	0.35	0.42	0.48	0.47	0.42
		Full Bayesian	0.45	0.45	0.49	0.39	0.30
		Simplified Bayesian	1.00	1.00	1.00	1.00	1.00

Across all configurations and sample sizes, the one-stage method (Naive) produces the highest bias and MSE, with severely poor coverage. The full Bayesian method (Full Bayesian) performs better than Naive but suffers from inflated bias and MSE in many cases, especially with larger sample sizes. In contrast, our proposed method (Simplified Bayesian) maintains low bias and MSE across all identification regimes and consistently achieves or exceeds nominal 95% coverage levels. Its advantage is most evident when $n = 500$, where Full Bayesian's estimation error increases markedly, while Simplified Bayesian remains stable.

2.4.3 Discussion of Results

The simulation findings highlight the practical advantages of our simplified Bayesian approach. Despite its reduced complexity—drawing Γ conditional only on (X, Z) — Simplified Bayesian achieves competitive accuracy and superior coverage relative to the full Bayesian method.

The strength of Simplified Bayesian lies in its modular estimation structure. By separating first-stage inference from the joint posterior, Simplified Bayesian reduces dependence among parameters, improving mixing and numerical stability in posterior sampling. In finite samples, this leads to more reliable interval estimates, especially when identification is weak or instruments are moderately informative.

These findings suggest that Simplified Bayesian is better, making it a strong candidate for practical applications.

2.5 Real Data Application

2.5.1 Data Description

We illustrate the proposed Bayesian instrumental variable approach using a dietary dataset derived from the US National Health and Nutrition Examination Survey (NHANES), which is a long-running research survey conducted by the National Center for Health Statistics. The goal of this longitudinal survey study is to assess the health and nutritional status of both adults and children in the United States, tracking the evolution of this status over time. During the 2009-2010 survey period, participants were interviewed and asked to provide their demographic background as well as information about nutrition habits. Participants also undertook a series of health examinations. To assess the nutritional habits of participants, dietary data were collected using two 24-hour recall interviews wherein the participants self-reported the consumed amount for a series of food items during the 24 hours prior to each interview.

In this analysis, we focus on analysing the relationship between the body mass index (BMI), a widely used measure of body composition and health risk, and three long-term energy, protein and fat consumption. For the ith participant, let y_i be their BMI, and $\mathbf{W}_i$ be a 3×1 vector of long-term energy, protein, and fat consumption, so we specify the outcome model to be

$$y_i = \mathbf{W}_i^\top \boldsymbol{\beta} + \tilde{\varepsilon}_i, \tag{2.5.1}$$

with $\tilde{\varepsilon}_i \sim N(0, \hat{\sigma})^2$ is the regression error for $i = 1, \ldots, n$. Nevertheless, this long-term consumption is not directly observed; the recorded 24-hour recalls should be only treated as surrogates (proxies) for it. Specifically, letting $\mathbf{D}_{ij}$ be the recorded amount from the jth 24-hour recall interviews, for $j = 1, 2$, a common model for such $\mathbf{D}_{ij}$ is

$$\mathbf{D}_{ij} = \mathbf{W}_i + \mathbf{V}_{ij},$$

where $\mathbf{V}_{ij}$ is an additive measurement error with mean zero independent from $\mathbf{W}_i$. To build a connection between the measurement error model with endogeneity, for any $j = 1, 2$, we can substitute $\mathbf{W}_i = \mathbf{D}_{ij} - \mathbf{V}_{ij}$ to the outcome model (2.5.1) to obtain

$$y_i = \mathbf{D}_{ij}^\top \boldsymbol{\beta} + \varepsilon_{ij}$$

where $\varepsilon_{ij} = \tilde{\varepsilon}_i - \mathbf{V}_{ij}^\top \boldsymbol{\beta}$. Here endogeneity arises since $\mathbf{D}_{ij}$ is correlated with ε_{ij} (through $\mathbf{V}_{ij}$), so ignoring measurement errors and run a linear model of the outcome on the self-recorded nutrition components on any day $\mathbf{D}_{ij}$ leads to biased estimates for $\boldsymbol{\beta}$; the same issue arises if we run the naive regression of y_i on the average $\mathbf{D}_i = (1/2)(\mathbf{D}_{i1} + \mathbf{D}_{i2})$.

To correct for measurement errors, we will treat $\mathbf{D}_{i1}$ as our endogenous covariates, i.e. $\mathbf{X}_i = \mathbf{D}_{i1}$ corresponding nutrient measures from day 2 as instruments, i.e. $\mathbf{Z}_i = \mathbf{D}_{i2}$. This IV approach to correct for measurement error is introduced in Carroll et al. (2006, Chapter 7). Letting $\mathbf{X}$ and $\mathbf{Z}$ be a $n \times 3$ matrix whose ith row is $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively, the linear IV model $\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{U}$ is exact when $\mathbf{W}_i$ and $\mathbf{V}_{ij}$ follow a multivariate Gaussian distribution; otherwise, we assumed this to be held approximately.

2.5.2 Parameter Estimates Comparison

We compare the coefficient estimates obtained from the naive one-stage Bayesian regression (Naive), which treats $\mathbf{X}$ as exogenous, with those from our proposed method (Simplified Bayesian), which accounts for endogeneity through instrumental variables.

Table 2.5.1 reports the posterior means and 95% credible intervals for each nutrient under both methods, and Figure 2.5.1 visualizes the posterior means and intervals, highlighting the contrast between Naive and Simplified Bayesian. The diamond-shaped points represent Naive, and the circles represent Simplified Bayesian, with horizontal bars showing 95% credible intervals. The results indicate substantial differences between the two approaches. For instance, the effect of energy intake on BMI is negative under both models, but the estimated magnitude is notably stronger under Simplified Bayesian, with a credible interval that excludes zero. This suggests that failing to account for endogeneity may underestimate the true association. Protein and fat show positive associations with BMI under both models, though the credible intervals under Naive are narrower and tend to be closer to zero. Given that the coverage rate for the Naive model was only 65% in simulation, this suggests the true effect of fat may not be well captured.

These differences underscore the importance of addressing endogeneity in observational dietary data. By incorporating valid instruments, Simplified Bayesian provides more reliable uncertainty quantification and potentially more accurate inference on the relationship between nutrient intake and body composition.

Table 2.5.1: Posterior estimates for BMI model under Naive and Simplified Bayesian

Method	Variable	Posterior Mean	95% CI Lower	95% CI Upper
Naive	Energy	−0.190	−0.290	−0.091
	Protein	0.206	0.134	0.277
	Fat	0.094	0.004	0.188
Simplified Bayesian	Energy	−0.611	−0.914	−0.303
	Protein	0.672	0.397	0.942
	Fat	0.212	−0.109	0.544

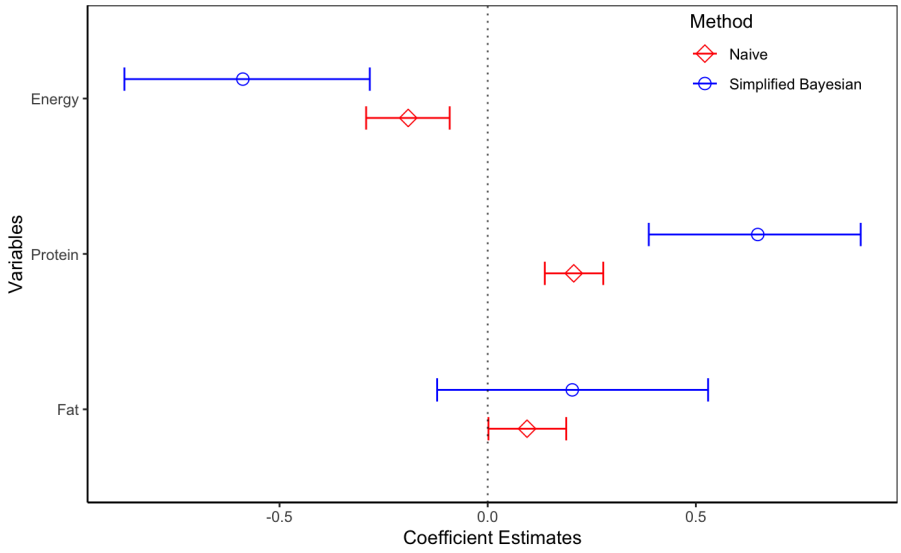

Figure 2.5.1: Posterior means and 95% credible intervals for the coefficients of Fat, Energy, and Protein model under the Naive and Simplified Bayesian approach in the NHANES study.

Chapter 3

Bayesian Lasso for Variable Selection in IV Models

3.1 Motivation

In many empirical applications, researchers face settings where the number of potential endogenous variables is large, so often only a subset of the included variables affect the outcome. Traditional instrumental variable methods, including our Full Bayesian and Simplified Bayesian approaches developed in Chapter 2, estimate all coefficients in $\boldsymbol{\beta}$ without considering that some may be exactly zero. This can lead to several problems, including inflated estimation uncertainty, reduced statistical power for detecting truly non-zero effects, and difficulties in interpretation when many small, potentially spurious coefficients are included (Hansen et al., 2008).

Therefore, variable selection in the IV context serves a dual purpose. First, it addresses the standard dimensionality concerns present in any regression setting by identifying the most relevant predictors. Second, and perhaps more critically, it helps distinguish between endogenous variables that genuinely affect the outcome and those that appear important merely due to their correlation with the error term. Without proper variable selection, we may incorrectly attribute causal significance to variables that are endogenous but not causally relevant.

As one of the most popular approach for variable selection, the Lasso estimator provides an attractive solution to simultaneously perform estimation and variable selection. In the frequentist framework, the Lasso estimator is obtained by augmenting the least square objective function with a ℓ_1 penalty, effectively shrinking all coefficients toward zero and setting others exactly to zero, hence performing automatic variable selection (Tibshirani, 1996). In the Bayesian framework, this penalty corresponds to placing independent Laplace priors on the regression coefficients, which allows for principled uncertainty quantification about both parameter values and variable inclusion (Park and Casella, 2008).

In this chapter, we will tackle the variable selection problem for the linear IV model using the Bayesian Lasso approach. This approach integrates seamlessly with our Simplified Bayesian framework developed in Section 2.3. Recall that Simplified Bayesian separates the estimation of first-stage coefficients $\boldsymbol{\Gamma}$ from the full joint posterior while maintaining the essential endogeneity correction through the joint modeling of error terms. This separation proves particularly advantageous when incorporating Lasso priors, as it allows the variable selection to focus on the second-

stage coefficients $\boldsymbol{\beta}$ without complicating the already simplified estimation of $\boldsymbol{\Gamma}$.

In our Bayesian Lasso extension of the Simplified Bayesian IV approach, we modify the prior specification for $\boldsymbol{\beta}$, replacing the standard normal prior with the hierarchical Lasso prior structure. The full conditional distributions for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ remain unchanged from the Simplified Bayesian framework established in equations (**??**) and (2.2.15), preserving both the computational advantages and the endogeneity correction mechanism. This approach ensures that variable selection operates on the relevant coefficients while maintaining the statistical properties that make Simplified Bayesian superior to Full in terms of computational efficiency and convergence.

3.2 Review of Lasso and Bayesian Lasso

The Least Absolute Shrinkage and Selection Operator (Lasso) introduces variable selection into regression by adding an ℓ_1 penalty to the standard least squares objective function. It is originally developed for regression models with exogenous covariates. In the presence of endogeneity, applying Lasso directly can lead to biased estimates. In the frequentist framework, Lasso estimates are obtained by solving:

$$\min_{\boldsymbol{\beta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\lambda \geq 0$ is the regularization parameter that controls the strength of the penalty. The key insight is that the ℓ_1 penalty $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$ has a unique geometric property: its constraint region has corners at the coordinate axes, which encourages solutions where some coefficients are exactly zero. This property distinguishes Lasso from ridge regression, which uses an ℓ_2 penalty and shrinks coefficients toward zero but rarely sets them exactly to zero. The ℓ_1 penalty's ability to produce sparse solutions makes it particularly valuable for variable selection, as it provides a principled way to automatically determine which variables should be included in the model (Tibshirani, 1996).

From a Bayesian perspective, Park and Casella (2008) demonstrated that Lasso estimates can be interpreted as the posterior model of estimates when the regression coefficients have independent Laplace (double exponential) priors. Specifically, if we place independent priors of the form:

$$p(\beta_j|\sigma^2)=\frac{\lambda}{2\sqrt{\sigma^2}}\exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right)$$

on each coefficient β_j, then the posterior mode corresponds exactly to the Lasso estimate with penalty parameter λ. This Bayesian interpretation provides several advantages over the frequentist Lasso. First, it yields full posterior distributions for all parameters rather than point estimates, enabling principled uncertainty quantification. Second, it provides a natural framework for hyperparameter selection through hierarchical modeling or empirical Bayes methods. Third, it facilitates the incorporation of prior information and the extension to more complex model structures, such as our instrumental variable setting. Park and Casella (2008) noted that the conditioning on σ^2 in the prior specification is crucial for ensuring a unimodal

posterior distribution. Without this conditioning, the posterior may exhibit multiple modes, which can lead to convergence difficulties in MCMC sampling and make point estimates less meaningful.

In the context of instrumental variable models, the Lasso penalty addresses a particular challenge: when many potential endogenous variables are available, standard IV methods may include variables that appear statistically significant due to their correlation with the error term rather than genuine causal effects. The Lasso penalty helps distinguish between these cases by shrinking the coefficients of variables with weak causal effects to zero.

3.3 Bayesian Lasso for IV Model

3.3.1 Model Specification

Building on the Simplified Bayesian framework, we next specify the complete hierarchical model for the Bayesian Lasso IV estimator. Recall that in our setting, our data consists of a $n \times 1$ vector $\mathbf{Y}$ response, a $n \times p$ matrix $\mathbf{X}$ of covariate, and a $n \times q$ matrix $\mathbf{Z}$ of instruments, which we assume to follow the model (1.2.1). The main parameter of interest is $\boldsymbol{\beta}$ that relates the effects of endogeneous covariables on the outcome.

Compared to the Simplified Bayesian approach developed in section 2.1.3, the key modification involves replacing the standard normal prior on $\boldsymbol{\beta}$ with the hierarchical Lasso structure, while retaining the same specifications for the first-stage coefficients and covariance matrix. The complete hierarchical specification of our Bayesian Lasso for the linear IV model is given as follows

$$
\begin{aligned}
&\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}, \sigma_\varepsilon^2 \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{I}_n) \\
&\mathbf{X}|\mathbf{Z}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_u \sim \text{Matrix-Normal}(\mathbf{Z}\boldsymbol{\Gamma}, \mathbf{I}_n, \boldsymbol{\Sigma}_u) \\
&\boldsymbol{\beta}|\tau_1^2, \ldots, \tau_p^2 \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{D}_\tau) \qquad (3.3.1) \\
&\mathbf{D}_\tau = \text{diag}(\tau_1^2, \ldots, \tau_p^2), \ \tau_j^2|\lambda^2 \sim \text{Exponential}(\lambda^2/2), \quad j = 1, \ldots, p \qquad (3.3.2) \\
&\boldsymbol{\Gamma} \sim \mathcal{N}(\boldsymbol{\mu}_\Gamma, \sigma_\Gamma^2 \mathbf{I}_{q \times p}) \\
&\boldsymbol{\Sigma} \sim \text{Inverse-Wishart}(\nu_0, \boldsymbol{\Psi}_0)
\end{aligned}
$$

The innovation lies in the specification of the prior for $\boldsymbol{\beta}$ in equation (3.3.1). Note that unlike the original Park and Casella (2008) formulation, we specify the prior for $\boldsymbol{\beta}$ without conditioning on σ_ε^2 to simplify the computational implementation. Rather than assuming a fixed covariance matrix, we introduce individual variance parameters τ_j^2 for each coefficient, collected in the diagonal matrix D_τ. These variance parameters are themselves random, following exponential distributions as specified in equation (3.3.2) with rate parameter $\lambda^2/2$. We note that, if we integrate out τ_j^2, the conditional priors for $\boldsymbol{\beta}$ in (3.3.1) and (3.3.2) is equivalent to the Laplace marginal prior for each coefficient,

$$\beta_j \sim \text{Laplace}(0, \lambda^2),$$

for $j = 1, \ldots, p$. However, by retaining the τ_j^2 parameters in our MCMC scheme, we can work with normal and exponential distributions throughout, which greatly simplifies the computational implementation. This representation has the additional

benefit of providing a natural interpretation for the individual variance parameters τ_j^2. Large values of τ_j^2 correspond to coefficients that the data suggest are non-zero, while τ_j^2 values shrunk toward zero correspond to coefficients that are likely to be exactly zero. For the regularization parameter λ^2, we place a Gamma prior with

$$\lambda^2 \sim \text{Gamma}(r, \delta) \tag{3.3.3}$$

where r and δ are hyperparameters that can be chosen to reflect prior beliefs about the degree of sparsity expected in $\boldsymbol{\beta}$.

3.4 Posterior Computation

Building on our Simplified Bayesian framework, most full conditional distributions remain unchanged, with modifications only for the parameters directly involved in the Lasso prior structure.

3.4.1 Joint Posterior Distributions

Based on the (2.2.2), and the prior distributions specified in Section 2.1, we can formulate the full joint posterior distribution under the Bayesian Lasso framework with endogenous covariates.

By Bayes' theorem, the posterior distribution is proportional to the product of the likelihood and the prior distributions:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \tau_1^2, \ldots, \tau_p^2, \lambda \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}) \propto\ & p(\mathbf{Y}, \mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \\ & \times p(\boldsymbol{\beta} \mid \tau_1^2, \ldots, \tau_p^2) \\ & \times p(\tau_1^2, \ldots, \tau_p^2 \mid \lambda) \\ & \times p(\boldsymbol{\Gamma}) \times p(\boldsymbol{\Sigma}) \times p(\lambda^2) \end{aligned} \tag{3.4.1}$$

Simplified Bayesian Lasso model assumes the following prior distributions for the parameters in the hierarchical structure:

$$\begin{aligned} p(\boldsymbol{\beta} \mid \tau_1^2, \ldots, \tau_p^2) &= (2\pi)^{-p/2} |D_\tau|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\top D_\tau^{-1} \boldsymbol{\beta}\right), \\ p(\tau_1^2, \ldots, \tau_p^2 \mid \lambda^2) &= \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau_j^2\right), \\ p(\boldsymbol{\Gamma}) &= (2\pi\sigma_\Gamma^2)^{-qp/2} \exp\left(-\frac{1}{2\sigma_\Gamma^2} \text{tr}\left[(\boldsymbol{\Gamma} - \boldsymbol{\mu}_\Gamma)^\top (\boldsymbol{\Gamma} - \boldsymbol{\mu}_\Gamma)\right]\right), \\ p(\boldsymbol{\Sigma}) &= |\boldsymbol{\Psi}_0|^{\nu_0/2} |\boldsymbol{\Sigma}|^{-(\nu_0+p+2)/2} \exp\left(-\frac{1}{2}\text{tr}(\boldsymbol{\Psi}_0 \boldsymbol{\Sigma}^{-1})\right), \\ p(\lambda^2) &= \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp\left(-\delta\lambda^2\right). \end{aligned}$$

3.4.2 Full Conditional Distribution for β

To derive the full conditional distribution for $\boldsymbol{\beta}$, we extract all terms from (3.4.1) that involve $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \tau_1^2, \ldots, \tau_p^2, \lambda^2) \propto p(\mathbf{Y}, \mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \times p(\boldsymbol{\beta} \mid \tau_1^2, \ldots, \tau_p^2)$$

The full conditional distribution for $\boldsymbol{\beta}$ retains the same form as derived in equation (2.2.6) for Simplified Bayesian, but with the covariance matrix now depending on the individual variance parameters:

$$\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \tau_1^2, \ldots, \tau_p^2, \lambda^2 \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_\beta, \tilde{\boldsymbol{\Sigma}}_\beta)$$

where:

$$\tilde{\boldsymbol{\Sigma}}_\beta = \left(\frac{1}{\sigma^2_{\varepsilon|u}} \mathbf{X}^T \mathbf{X} + D_\tau^{-1} \right)^{-1}$$
$$\tilde{\boldsymbol{\mu}}_\beta = \tilde{\boldsymbol{\Sigma}}_\beta \frac{1}{\sigma^2_{\varepsilon|u}} \mathbf{X}^T \mathbf{W}$$

with $\sigma^2_{\varepsilon|u}$ and $\mathbf{W}$ defined as in equation (2.2.4), and $D_\tau^{-1} = \text{diag}(1/\tau_1^2, \ldots, 1/\tau_p^2)$.

3.4.3 Full Conditional Distribution for τ_j^2

To derive the full conditional distribution for τ_j^2, we extract all terms from (3.4.1) that involve τ_j^2:

$$p(\tau_j^2 \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \lambda^2) \propto p(\boldsymbol{\beta} \mid \tau_1^2, \ldots, \tau_p^2) \times p(\tau_j^2 \mid \lambda^2)$$

Since these terms do not involve the endogeneity structure, the derivation of the full conditional distribution for τ_j^2 remains unchanged. It is proportional to the product of the Gaussian prior on $\boldsymbol{\beta}$ given τ_j^2, and the exponential prior on τ_j^2 given λ^2:

$$p(\tau_j^2 \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \lambda^2) \propto \frac{1}{\sqrt{\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right) \cdot \exp\left(-\frac{\lambda^2 \tau_j^2}{2}\right)$$
$$\propto (\tau_j^2)^{-1/2} \exp\left(-\frac{\beta_j^2}{2\tau_j^2} - \frac{\lambda^2 \tau_j^2}{2}\right)$$

Letting $\eta_j = 1/\tau_j^2$, we obtain the posterior distribution of η_j in closed form to be

$$\frac{1}{\tau_j^2} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \lambda^2 \sim \text{Inverse-Gaussian}\left(\mu' = \frac{\lambda}{|\beta_j|},\ \lambda' = \lambda^2\right)$$

3.4.4 Full Conditional Distribution for λ^2

To derive the full conditional distribution for λ^2, we extract all terms from (3.4.1) that involve λ^2:

$$p(\lambda^2 \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \tau_1^2, \ldots, \tau_p^2) \propto p(\tau_1^2, \ldots, \tau_p^2 \mid \lambda^2) \cdot p(\lambda^2)$$

Since the terms involving λ^2 do not depend on the endogeneity structure, the derivation of its full conditional distribution is unchanged. It is proportional to

the product of the exponential prior on τ_j^2 given λ, and the Gamma prior on λ^2. Substituting the expressions for each component, we have:

$$p(\lambda^2 \mid \cdots) \propto \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2 \tau_j^2}{2}\right) \cdot \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp(-\delta\lambda^2)$$
$$\propto (\lambda^2)^{p+r-1} \exp\left(-\lambda^2 \left(\frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + \delta\right)\right)$$

This is the kernel of a Gamma distribution. Therefore, the full conditional distribution is given by

$$\lambda^2 \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \tau_1^2, \ldots, \tau_p^2 \sim \text{Gamma}\left(p + r,\ \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + \delta\right).$$

3.4.5 Full Conditional Distributions for Γ and Σ

The full conditional distributions for the first-stage coefficients $\boldsymbol{\Gamma}$ and the covariance matrix $\boldsymbol{\Sigma}$ remain exactly as specified in the Simplified Bayesian framework. For $\boldsymbol{\Gamma}$, we continue to use the simplified full conditional distribution given in equations (**??**) and (**??**), while for $\boldsymbol{\Sigma}$, we use the full conditional distribution specified in equation (2.2.15).

This preservation ensures that the endogeneity correction mechanism is maintained while the variable selection operates specifically on the second-stage coefficients.

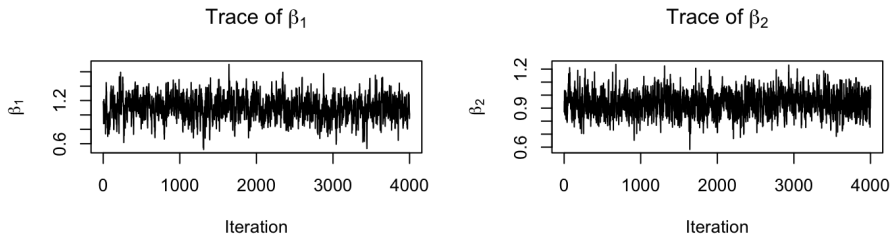

Figure 3.4.1: Trace plots of β_1 and β_2 under the Bayesian Lasso posterior. The chains exhibit good mixing and convergence behavior.

3.5 Simulation Studies

3.5.1 Simulation Setup

To evaluate the performance of the Bayesian Lasso IV estimator in sparse settings, we design a simulation in which only a subset of the covariates have nonzero effects. Specifically, we set $p = 10$ and $q = 10$, with a sample size of $n = 500$. The true coefficient vector is $\boldsymbol{\beta}_{\text{true}} = (1.5, -0.5, 0.8, 0, 0, 0, 0, 0, 0, 0)^T$, where only the first three variables have a true effect on the outcome. The instrument matrix $\mathbf{Z}$ has each element generated from the standard normal, and the endogenous covariates $\mathbf{X}$ are constructed via a first-stage model $\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{U}$, where each element of $\boldsymbol{\Gamma}$ is

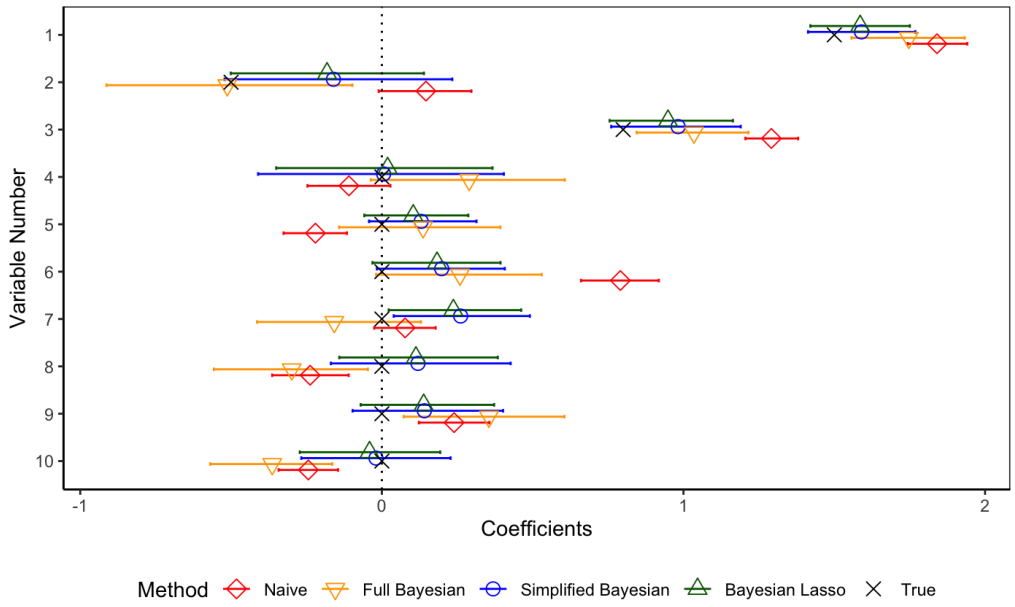

Figure 3.5.1: Posterior means and 95% credible intervals for each coefficient in a single replication. The true values are marked with black crosses.

drawn from a uniform distribution on $(-0.5, 0.5)$ and $(\mathbf{U}, \boldsymbol{\varepsilon})$ are jointly drawn from a multivariate normal distribution with positive-definite covariance matrix. The response $\mathbf{Y}$ is generated as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_{\text{true}} + \boldsymbol{\varepsilon}$.

We estimate the model using four approaches: Naive, Full Bayesian, Simplified Bayesian and Bayesian Lasso. For the Bayesian Lasso estimator, we initialized the shrinkage parameter at $\lambda = 0.5$, corresponding to a moderately strong regularization prior. The parameter was subsequently updated during the MCMC sampling process. Each method is applied to 100 independent simulated datasets. For each run, we record the posterior mean 95% credible intervals, and variable selection performance on the coefficient vector $\boldsymbol{\beta}$.

To complement the simulation, we also visualize the posterior coefficient estimates and credible intervals from a single simulated dataset. As shown in Figure 3.5.1, the plot compares the estimated coefficients across all four methods. The true values of the coefficients are marked with black crosses. We observe that Bayesian Lasso produces tighter intervals for zero coefficients and less shrinkage for relevant variables compared to Naive and Full Bayesian.

3.5.2 Variable Selection Performance

We assess variable selection performance by checking whether the 95% posterior credible interval for each coefficient β_j excludes zero. A coefficient is considered selected if zero is not included in its posterior interval. Based on this rule, we compute:

- **TP** (True Positives): Nonzero coefficients correctly identified as nonzero.
- **FP** (False Positives): Zero coefficients incorrectly selected as nonzero.
- **FPR** (False Positive Rate): FP divided by the number of true zero coefficients.

- **FNR** (False Negative Rate): FN divided by the number of true nonzero coefficients.
- **Precision**: TP divided by the total number of selected coefficients.

These metrics evaluate how well each method detects relevant variables while avoiding the inclusion of irrelevant ones.

Each method is run on 100 simulated datasets generated under the same sparse data-generating process, as described in Section 3.5.1. The sample size is set to $n = 500$, with $p = q = 10$. The true coefficient vector includes only three nonzero entries. For the Bayesian Lasso estimator (Bayesian Lasso), We initialized the shrinkage parameter at $\lambda = 1$, corresponding to the default starting value in our implementation. The parameter was subsequently updated within the MCMC chain. All methods were implemented using 2000 total MCMC iterations with a burn-in of 500.

Table 3.5.1 reports the overall performance metrics averaged over 100 replications. Among the methods, Bayesian Lasso achieves the lowest MSE (0.4942), the highest coverage rate (0.5990), and the best precision (0.5603). Notably, both Simplified Bayesian and Bayesian Lasso yield near-zero false negative rates (0.0067), confirming their ability to recover the true nonzero coefficients. Compared to Naive and Full Bayesian, which suffer from high false positive rates and low coverage, the instrumental variable approaches Simplified Bayesian and Bayesian Lasso perform substantially better across all criteria.

The relatively moderate precision values across all methods (e.g., 0.5603 for Bayesian Lasso) may partially reflect the conservative nature of using 95% credible intervals for variable selection, particularly under limited iterations. Since the simulations used 2000 iterations with 500 burn-in, longer chains might improve interval stability and reduce false positives further.

Table 3.5.1: Overall performance metrics across 100 simulations

Method	MSE	Coverage	FPR	FNR	Precision	TP
Naive	0.9978	0.1740	0.8171	0.0333	0.3412	2.97
Full Bayesian	0.9379	0.3900	0.6071	0.0467	0.4248	2.95
Simplified Bayesian	0.5189	0.5890	0.4129	0.0067	0.5542	2.99
Bayesian Lasso	0.4942	0.5990	0.4029	0.0067	0.5603	2.99

In addition to overall performance, we report the average estimation bias for each coefficient across the four methods. As shown in Table 3.5.2, Bayesian Lasso consistently produces low bias estimates for both active and inactive variables. For zero coefficients (β_4 to β_{10}), the average bias is centered near zero, indicating effective shrinkage. For the nonzero coefficients (β_1 to β_3), all four methods produce reasonably small bias, with Simplified Bayesian and Bayesian Lasso performing best overall.

Table 3.5.2: Average bias by coefficient across 100 simulations

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Naive	-0.027	0.020	-0.027	0.058	-0.048	-0.003	-0.018	0.049	-0.050	0.024
Full Bayesian	0.086	-0.030	0.043	0.051	-0.055	-0.010	-0.029	0.045	-0.041	0.033
Simplified Bayesian	-0.010	-0.010	-0.030	0.019	-0.038	-0.013	0.019	0.002	-0.023	0.013
Bayesian Lasso	-0.018	-0.003	-0.031	0.020	-0.038	-0.010	0.019	0.002	-0.024	0.016

Overall, the Bayesian Lasso approach (Bayesian Lasso) delivers a favorable trade-off between variable selection accuracy and estimation efficiency. It preserves the endogeneity correction structure of Simplified Bayesian, while leveraging shrinkage priors to improve sparsity control, coverage, and bias.

3.6 Real Data Application

We apply the proposed methods to a dataset from the same NHANES study in Section 2.5. However, instead of examining the relatioship between BMI (outcome) on three total nutrition variables, we now study its relationship with $p = 42$ nutritional daily intakes, ranging from such as sugars, total vitamins, retinol, lycopene, zinc, and selenium, among many others. Specifically, we selected the endogeneous covariates $\mathbf{X}$ are derived from nutrient intake on Day 1, while the instruments $\mathbf{Z}$ are constructed using the corresponding variables from Day 2. All covariates and instruments were standardized prior to estimation to account for different scales among these variables. No missing values were present in the subset of data used for analysis.

We estimated the models using three approaches: Naive, Simplified Bayesian, and Bayesian Lasso. For the Bayesian Lasso, we set the shrinkage parameter to $\lambda = 1$, using the default setting in the algorithm. Each method was run using 1000 MCMC iterations with a burn-in of 200. A larger number of iterations may lead to more precise posterior estimates, albeit with increased computational cost.

Figure 3.6.1 displays the variables identified as significant under each method. A variable is considered significant if its 95% credible interval excludes zero. The corresponding posterior estimates and credible intervals are summarized in Table 5.0.1 in the Appendix. The figure plots the posterior means and corresponding intervals for all selected variables.

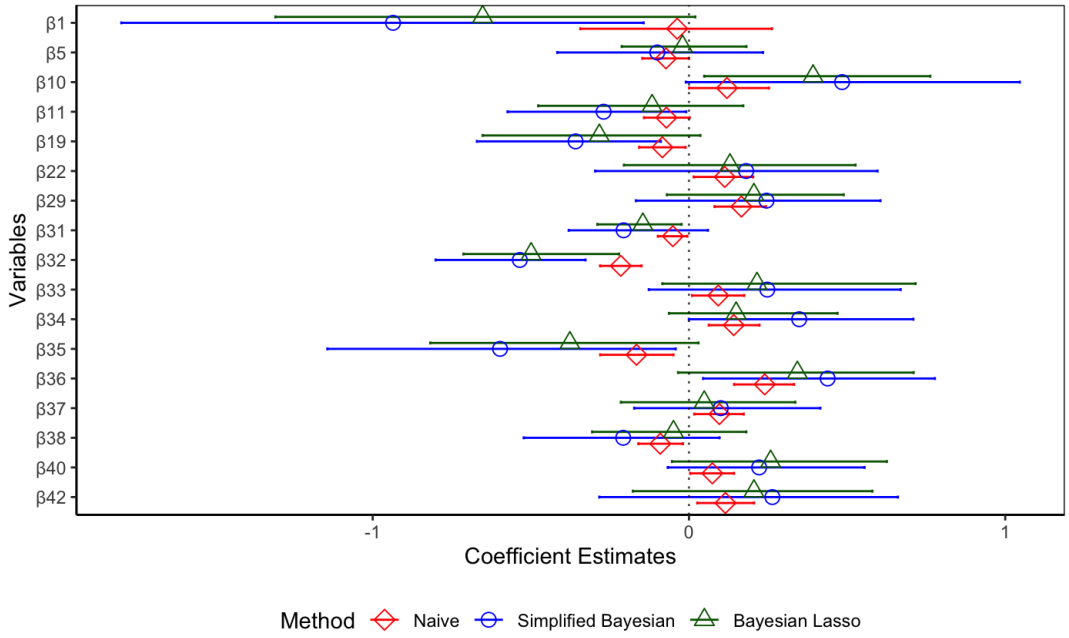

Figure 3.6.1: Posterior coefficient estimates and 95% credible intervals for selected variables. Only variables with intervals excluding zero are shown.

Naive identifies 15 significant variables, including many that are likely correlated with BMI through confounding. In contrast, Simplified Bayesian selects only 6 variables, among which β_1 (Energy), β_{11} (Vitamin E as alpha-tocophero), and β_{19} (Beta-carotene) are consistent with Naive.

Bayesian Lasso identifies only 3 variables: β_{10} (Cholesterol), β_{31} (Added Vitamin B12), and β_{32} (Vitamin C). All three have relatively tight credible intervals excluding zero, and were also selected by Naive. Notably, β_{32} (Vitamin C) is selected across all three methods, suggesting it may be a robust predictor of BMI. Compared to Simplified Bayesian, the Lasso prior in Bayesian Lasso leads to more aggressive shrinkage of coefficients and clearer variable selection.

Table 3.6.1 summarizes the set of significant variables selected by each method.

Table 3.6.1: Number and identity of variables with credible intervals excluding zero

Method	Variable	Nutrient Name
Naive	β_5	Dietary fiber (gm)
	β_{10}	Cholesterol (mg)
	β_{19}	Beta-carotene (mcg)
	β_{22}	Lutein + zeaxanthin (mcg)
	β_{29}	Vitamin B12 (mcg)
	β_{31}	Added Vitamin B12 (mcg)
	β_{32}	Vitamin C (mg)
	β_{33}	Vitamin D (mg)
	β_{34}	Vitamin K (mg)
	β_{35}	Calcium (mg)
	β_{36}	Phosphorus (mg)
	β_{37}	Magnesium (mg)
	β_{38}	Iron (mg)
	β_{40}	Copper (mg)
	β_{42}	Potassium (mg)
Simplified Bayesian	β_1	Energy (kcal)
	β_{11}	Vitamin E as alpha-tocopherol (mg)
	β_{19}	Beta-carotene (mcg)
	β_{32}	Vitamin C (mg)
	β_{35}	Calcium (mg)
	β_{36}	Phosphorus (mg)
Bayesian Lasso	β_{10}	Cholesterol (mg)
	β_{31}	Added Vitamin B12 (mcg)
	β_{32}	Vitamin C (mg)

These results illustrate the regularization advantage of the Bayesian Lasso approach. While the naïve method identifies many variables likely due to omitted variable bias or endogenous confounding, the Lasso-based IV model focuses on a much sparser subset of dietary variables with strong and stable associations with BMI.

Chapter 4

Conclusion

4.1 Summary of Key Findings

This thesis developed and evaluated Bayesian approaches for addressing endogeneity and variable selection in instrumental variable models. The research tackled two fundamental challenges that frequently arise in empirical analysis: correcting for endogeneity bias and performing variable selection in sparse settings where only a subset of potential predictors are truly relevant.

The key empirical findings demonstrate that the Simplified Bayesian approach consistently outperforms both naive regression and the Full Bayesian method across different sample sizes and identification regimes. In comprehensive simulation studies spanning various configurations of instruments and endogenous variables, Simplified Bayesian achieved superior bias reduction, mean squared error performance, and coverage rates while maintaining computational efficiency.

The Bayesian Lasso extension proved particularly effective for sparse settings, successfully identifying relevant variables while shrinking irrelevant coefficients toward zero. This approach addresses a critical gap in existing literature by combining endogeneity correction with principled variable selection. In applications to NHANES dietary data involving 42 nutritional variables, the method identified a parsimonious set of three key factors associated with BMI, demonstrating practical utility for sparse settings where researchers face many potential predictors but seek interpretable models. The contrast with naive methods, which selected 15 variables likely due to confounding, highlights the importance of addressing endogeneity in variable selection procedures.

4.2 Limitations

The methods assume linear relationships and Gaussian error structures, which may not hold universally. Prior specifications, particularly for regularization parameters, can influence results. The simulation studies focused on moderate sample sizes and may not capture performance with very small samples or extremely weak instruments.

A notable limitation is the unexpectedly poor performance of the Full Bayesian approach. Despite theoretical appeal, it exhibited higher bias and deteriorating performance with larger samples. The underlying reasons remain unclear and require further investigation.

4.3 Future Research Directions

Future extensions could address non-linear models, discrete endogenous variables, and automatic prior selection procedures. Investigating robustness to normality violations and panel data applications would enhance practical utility. Understanding why the Full Bayesian approach underperforms is a critical research priority. This could involve analyzing posterior dependencies, exploring alternative priors, or investigating different MCMC strategies to determine when joint versus modular estimation is preferable.

Chapter 5

Appendix

Table 5.0.1: Posterior estimates and 95% credible intervals for variables identified as significant

Method	Variable	Estimate	95% CI Lower	95% CI Upper
Naive	β_5	-0.0727	-0.1475	-0.00004
Naive	β_{10}	0.1202	0.0003	0.2528
Naive	β_{19}	-0.0843	-0.1576	-0.0110
Naive	β_{22}	0.1132	0.0148	0.2301
Naive	β_{29}	0.1657	0.0808	0.2432
Naive	β_{31}	-0.0510	-0.0984	-0.0005
Naive	β_{32}	-0.2157	-0.2811	-0.1499
Naive	β_{33}	0.0925	0.0096	0.1748
Naive	β_{34}	0.1415	0.0624	0.2228
Naive	β_{35}	-0.1660	-0.2808	-0.0491
Naive	β_{36}	0.2395	0.1426	0.3322
Naive	β_{37}	0.0961	0.0167	0.1731
Naive	β_{38}	-0.0908	-0.1603	-0.0190
Naive	β_{40}	0.0739	0.0038	0.1424
Naive	β_{42}	0.1152	0.0263	0.2059
Simplified Bayesian	β_1	-0.9358	-1.7948	-0.1432
Simplified Bayesian	β_{11}	-0.2699	-0.5737	-0.0089
Simplified Bayesian	β_{19}	-0.3586	-0.6706	-0.0889
Simplified Bayesian	β_{32}	-0.5348	-0.8011	-0.3272
Simplified Bayesian	β_{35}	-0.5973	-1.1434	-0.4194
Simplified Bayesian	β_{36}	0.4384	0.0446	0.7776
Bayesian Lasso	β_{10}	0.3925	0.0481	0.7631
Bayesian Lasso	β_{31}	-0.1457	-0.2897	-0.0233
Bayesian Lasso	β_{32}	-0.4992	-0.7131	-0.2209

Bibliography

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Hansen, C., Hausman, J., and Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422.

Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., and Walter, S. L. (2021). Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management*, 47(1):105–143.

Lin, W., Feng, R., and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the american statistical association*, 103(482):681–686.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.