Bayesian Estimation with Flexible Priors for the Instrumental Variabls Models

Julianne Swenson* John Hsu†

Abstract

The instrumental variables method is a valuable tool in the analysis of simultaneous equations models. Since the estimation of coefficients in the model can be challenging, adept modeling of the covariance matrix is also important. The Inverse Wishart distribution is commonly used to provide a conjugate prior for the covariance matrix. However, the Inverse Wishart is limited in the flexibility to model prior information. We propose an alternative that allows for the specification of varying confidence levels for each element in the covariance matrix for the Instrumental Variables models.

Key Words: covariance matrix, Inverted Wishart prior, Markov Chain Monte Carlo, matrix logarithm, Metropolis-Hastings algorithm

1. Introduction

Improving estimation techniques for the error term in simultaneous equations models (SEM) presents a unique opportunity to make a more informed determination of the existence of causality in a relationship of interest. SEMs are very prevalent in the field of econometrics – an example of which is the ubiquitous supply and demand curves. As their name would suggest, they consist of multiple equations that interact to produce the observed data. This type of interaction implies the existence of endogenous variables – those that are jointly determined by the model. As a result, the approach of considering the equations individually is not a viable option, as a "simultaneity" bias will occur when an independent variable is correlated with the error term. This forces the dependent variables to be modeled simultaneously by the system of equations.

The method of instrumental variables (IV) provides a way of consistently estimating coefficients in SEMs. By deconstructing and rebuilding the simultaneous equations with instruments, a more insightful look into the inherent causal relationships is possible. In SEMs, variables are candidates for instruments if they are correlated with the endogenous explanatory variable (conditional on other covariates), but uncorrelated with the disturbance term in the explanatory equation. When these instrumental variables are combined in regression, it is often referred to as an IV regression model.

Research into IV regression began with classical/frequentist methods. In the late 1940s and early 1950s, the first inferential procedure was developed using the method of maximum likelihood (Anderson and Rubin, 1949; Koopmans and Hood, 1953). This method, called limited information maximum likelihood (LIML), derives the maximum likelihood estimator for the parameters of the reduced form model.

Soon after, Theil (1953) and Basmann (1957) developed an alternative inferential procedure called two stage least squares (2SLS). Unlike LIML, it does not require any distributional assumptions for independent variables on the right hand side of the equations. Instead, it utilizes two steps to determine the estimator: first, perform Ordinary Least Squares (OLS) regression of the endogenous variable on the instruments to obtain predictions for the endogenous variable; second, perform OLS regression of the dependent variable on the

*Department of Statistics and Applied Probability, University of Californian, Santa Barbara, CA 93106
†Department of Statistics and Applied Probability, University of Californian, Santa Barbara, CA 93106

predicted values from step one. To this day, 2SLS is the most commonly used estimation method as it is not only efficient (it takes all of the information contained in the set of instruments and instills it into a single instrument) but also simple to use. It is worth noting that, under certain circumstances, the LIML estimator is the same as the 2SLS estimator.

Interest in Bayesian approaches to IV regression sparked years later with a seminal publication by Drèze (1976). In his paper on the Bayesian analysis of SEMs, Drèze proposed to equalize the Bayesian and classical analysis of IV models by using sufficiently diffuse priors for the parameters of interest. Today, Drèze's method is considered by many to be a Bayesian version of 2SLS. Yet, as many have noted, its greatest flaw is that it does not consider the effect of weak instruments (those that are poor predictors of the endogenous variable) on the inference of IV models.

Additional Bayesian analyses followed in the research of Zellner (1971), Drèze and Morales (1976), Drèze and Richard (1983), and Tsurumi (1985, 1990). More recent research tackles the problem present in Drèze's method – that of weak instruments on the inference of IV models. The problem with weak instruments is that they can cause near non-identification of structural parameters. This issue is identified and addressed in papers such as Kleibergen (1997), Kleibergen and van Dijk (1997), Chao and Phillips (1998), and Kleibergen and Zivot (2003).

Until recently, Bayesian analysis of IV regression focused primarily on the estimation of parameters in the relationship of interest. While this allowed researchers to make reasonable conclusions about the direction of causality, these conclusions may not be as well supported as they would like to believe. Since choosing an appropriate instrument is often a difficult task, it is only natural that many chosen instruments only explain a small proportion of the variability in the endogenous variables. Therefore, accurate estimation of the error term could provide further insight on the relationship of interest and play a significant role in determining causality.

The current standard Bayesian method for estimation of the covariance matrix of the error term, $\mathbf{\Sigma}$, uses an Inverse-Wishart (IW) prior. Its popularity stems from the simplicity it offers as a natural conjugate prior. Although this conjugacy property does simplify posterior inference, it has several significant flaws. The main problem lies in the parameters of the IW distribution – one location matrix and one degree of freedom. While the location matrix is sufficient, the degree of freedom parameter is not sufficient to fully represent the confidence levels for all the elements of the location matrix.

We propose instead the specification of a flexible prior that addresses the flaws seen in using the IW prior. By considering its matrix logarithm transformation instead of $\mathbf{\Sigma}$ itself, we can specify a multivariate Normal for the vector of elements of the transformed covariance matrix. The beauty of the multivariate Normal prior is that we can specify locational information as well as *different* levels of confidence for each element. This allows researchers to fully utilize any information they may have, *a priori*, regarding the various elements of covariance matrix of the error term.

Due to the nature of our proposed method, direct sampling is not a viable option. Markov Chain Monte Carlo (MCMC) sampling is used instead to perform posterior inference. MCMC sampling allows the determination of the characteristics of a density when it is difficult to obtain that density analytically or numerically. It creates a sequence of random samples from the density, allowing for the calculation of characteristics such as the mean and variation. Our use of MCMC sampling will utilize the Metropolis-Hastings accept-reject algorithm.

The rest of this chapter is organized as follows. Section 2 introduces the IV regression model. It provides an overview of the different forms used for IV regression and a discussion on the benefits of each. Section 3 explores a few Bayesian methods currently used in

IV regression analysis. In Section 4, we introduce our method. We provide an in depth look at the math and logic for the procedure. Section 5 is used to conclude.

2. The Model

The basic two equation IV regression model seen in Rossi et al. (2005) is used for calculations, but can be extended to a more general multivariate case. In this model, y is the response variable and x_i is the endogenous regressor, for $i = 1, 2, \ldots, n$, where n is the sample size. The model is then defined by the following system of equations:

$$x_i = \mathbf{z}_i'\boldsymbol{\delta} + \epsilon_{1i} \tag{1}$$

$$y_i = \kappa + \beta x_i + \epsilon_{2i} \tag{2}$$

In equation (1), $\mathbf{z}_i$ is an h-dimensional vector containing both the intercept term and $h-1$ instruments, $\boldsymbol{\delta}$ is the corresponding h-dimensional vector of coefficients, and ϵ_{1i} is the error term. Note that when $\boldsymbol{\delta} = \mathbf{0}$, i.e. the instruments do not explain any variability in x, x is simply an error term and y is simply the sum of a constant and a disturbance term. Clearly, this limiting case would make the model unidentifiable.

In equation (2), κ is the intercept, β is the causal effect parameter, and ϵ_{2i} is the error term. It is assumed that the instruments in $\mathbf{z}_i$ are related to x_i, but uncorrelated with ϵ_{2i}. If we let,

$$\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \epsilon_{2i})'$$

it is assumed that $\boldsymbol{\epsilon}_i$ are i.i.d. $N_2(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a 2×2 positive definite matrix. In other words, the error vector $\boldsymbol{\epsilon}_i$ is distributed as a bivariate normal distribution with zero mean vector and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$ covariance matrix.

Although the error vector $\boldsymbol{\epsilon}_i$ is distributed as a bivariate normal distribution, this model is not that of a standard bivariate regression. The potential existence of a correlation between ϵ_{1i} and ϵ_{2i} translates to a potential corresponding correlation between x_i and ϵ_{2i}. If in existence, this correlation would create an "endogenity" bias when estimating the value of β. In other words, if x_i and ϵ_{2i} are correlated, then β cannot be consistently estimated since information about x_i that is correlated with ϵ_{2i} will be used in the estimation.

Structural form modeling is built upon the belief of a theoretical model as well as assumptions about structural errors. Since the goal of structural form modeling is to estimate the parameters of behavioral functions, it is very popular in econometrics. In the case of IV regression, the use of a structural form has the added benefit of allowing for endogenous variables in the system of equations. Here, this structural form system of equations can be simplified by defining the following:

$$\mathbf{Y}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 0 & \boldsymbol{\delta}^T \\ \kappa & \beta & \mathbf{0}^T \end{bmatrix} \quad \mathbf{X}_i = \begin{bmatrix} 1 \\ x_i \\ \mathbf{z}_i \end{bmatrix}.$$

This allows the structural form model to be expressed as:

$$\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \boldsymbol{\epsilon}_i \tag{3}$$

where $\mathbf{Y}_i$ is a (2×1) vector, $\mathbf{X}_i$ is $((h+2) \times 1)$ matrix, and $\mathbf{A}$ is a $(2 \times (h+2))$ matrix.

3. Privious Methods

3.1 Inverse Wishart Prior

An Inverse Wishart prior is commonly used for the covariance matrix $\mathbf{\Sigma}$ of the error term. From equation (3), the joint likelihood of $\mathbf{A}$ and $\mathbf{\Sigma}$ can be expressed as:

$$\begin{aligned}\ell(\mathbf{A},\mathbf{\Sigma}|\mathbf{Y}) &= (2\pi)^{-n}|\mathbf{\Sigma}|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i-\mathbf{A}\mathbf{X}_i)'\mathbf{\Sigma}^{-1}(\mathbf{Y}_i-\mathbf{A}\mathbf{X}_i)\right\}\\ &= (2\pi)^{-n}|\mathbf{\Sigma}|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}tr(\mathbf{W}\mathbf{\Sigma}^{-1})\right\} \qquad (4)\end{aligned}$$

where

$$\mathbf{W}=\sum_{i=1}^{n}(\mathbf{Y}_i-\mathbf{A}\mathbf{X}_i)(\mathbf{Y}_i-\mathbf{A}\mathbf{X}_i)'$$

such that $\frac{1}{n}\mathbf{W}$ is the maximum likelihood estimator for $\mathbf{\Sigma}$ and $\mathbf{Y}=(Y_1,\cdots,Y_n)'$. Now with the prior specification of $\mathbf{\Sigma}\sim IW(\nu_0,\mathbf{\Sigma}_0)$, the conditional posterior density of $\mathbf{\Sigma}$ is simply the product of the prior density and the profile likelihood function of $\mathbf{\Sigma}$:

$$\begin{aligned}\pi(\mathbf{\Sigma}|\mathbf{Y}) &\propto |\mathbf{\Sigma}|^{-\frac{v_0+3}{2}}\exp\left\{-\frac{1}{2}tr(\mathbf{\Sigma}_0\mathbf{\Sigma}^{-1})\right\}\times|\mathbf{\Sigma}|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}tr(\mathbf{W}\mathbf{\Sigma}^{-1})\right\}\\ &\propto |\mathbf{\Sigma}|^{-\frac{v_0+n+3}{2}}\exp\left\{-\frac{1}{2}tr[(\mathbf{\Sigma}_0+\mathbf{W})\mathbf{\Sigma}^{-1}]\right\} \qquad (5)\end{aligned}$$

The resulting conditional posterior is then conveniently also distributed as IW:

$$\mathbf{\Sigma}|\mathbf{Y}\sim IW(v_0+n,\mathbf{\Sigma}_0+\mathbf{W})$$

Since the IW prior is a natural conjugate prior specification for $\mathbf{\Sigma}$, it is not surprising that is so commonly used.

When the prior specifications and subsequent posterior distributions are determined, Gibbs sampling can be used to perform posterior inference. Rossi et al. (2005) suggested the following prior specification for the structural parameters:

$$\boldsymbol{\delta}\sim N_2(\mathbf{d}_0,\mathbf{D}_0),\qquad (\kappa,\beta)'\sim N_2(\mathbf{b}_0,\mathbf{B}_0),\qquad \mathbf{\Sigma}\sim IW(v_0,\mathbf{\Sigma}_0)$$

where the hyperparameters $\mathbf{d}_0,\mathbf{D}_0,\mathbf{b}_0,\mathbf{B}_0,v_0$, and $\mathbf{\Sigma}_0$ are specified. They reported that the Gibbs sampling can be performed as follows:

1. Sample $\mathbf{\Sigma}$. As seen above, the covariance matrix of the error term has the conditional posterior
$$\mathbf{\Sigma}|\mathbf{Y},\boldsymbol{\delta},\kappa,\beta\sim IW(v_0+n,\mathbf{\Sigma}_0+\mathbf{W})$$
2. Sample (κ,β). The regression parameters (κ,β) have a joint conditional posterior
$$\kappa,\beta|\mathbf{Y},\boldsymbol{\delta},\mathbf{\Sigma}\sim N_2(\mathbf{b}_1,\mathbf{B}_1)$$
where
$$\mathbf{B}_1=\left(\mathbf{B}_0^{-1}+\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i{}'\right)^{-1},\qquad \mathbf{b}_1=\mathbf{B}_1\left(\mathbf{B}_0^{-1}\mathbf{b}_0+\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\tilde{y}_i\right)$$
$$\tilde{\mathbf{x}}_i=\frac{1}{\sigma_{2|1}^{\frac{1}{2}}}(1,x_i)',\qquad \tilde{y}_i=\frac{1}{\sigma_{2|1}^{\frac{1}{2}}}\left[y_i-(x_i-\mathbf{z}_i'\boldsymbol{\delta})\frac{\sigma_{12}}{\sigma_{11}}\right],\qquad \sigma_{2|1}=\sigma_{22}(1-\rho^2)$$
and $\rho=\sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$.

3. Sample $\boldsymbol{\delta}$. The regression parameter $\boldsymbol{\delta}$ has the conditional posterior

$$\boldsymbol{\delta}|\mathbf{Y}, \kappa, \beta, \boldsymbol{\Sigma} \sim N_2(\mathbf{d}_1, \mathbf{D}_1)$$

where

$$\mathbf{D}_1 = \left(\mathbf{D}_0^{-1} + \sum_{i=1}^{n} \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i{}'\right)^{-1}, \qquad \mathbf{d}_1 = \mathbf{D}_1 \left(\mathbf{D}_0^{-1}\mathbf{d}_0 + \sum_{i=1}^{n} \tilde{\mathbf{z}}_i \tilde{x}_i\right)$$

$$\tilde{x}_i = \frac{1}{\sigma_{1|2}^{\frac{1}{2}}}\left[x_i - (y_i - \kappa - \beta x_i)\frac{\sigma_{12}}{\sigma_{22}}\right], \qquad \tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\sigma_{1|2}^{\frac{1}{2}}} \qquad \text{and} \qquad \sigma_{1|2} = \sigma_{11}(1-\rho^2)$$

This methodology illustrates that an IW prior for $\boldsymbol{\Sigma}$ is not only convenient for its natural conjugacy property, but also very easy to implement. Yet, as previously discussed, its inability to specify varying confidence levels for each of the elements of the covariance matrix makes it very limited in its ability to model information known *a priori*.

3.2 Dirichlet Process Prior

Conley et al. (2008) proposed an alternative Bayesian approach to IV regression. They assume the same structure for the IV regression, but model the error distributions non-parametrically. Since most researchers consider the assumption of normality of the error terms in IV models as only an approximation, this proposed method's flexible error distribution allows for a more accurate estimation of the true error distribution.

Conley et al.'s approach uses a Dirichlet Process (DP) prior for the error terms. This method strives to better capture the structure of the data. In particular, this approach allows the error terms to be modeled using the same distribution, but with varying parameters. This creates, in effect, a mixture model. Although the number of base distributions is not fixed *a priori*, the DP prior and the data will help determine the number of mixture components. This is accomplished by the process itself since it encourages the grouping of "similar" observations.

Similar to previous approaches (see Section 3.1), it specifies normal priors for the other parameters of interest $\boldsymbol{\delta} \sim N_2(\mathbf{d}_0, \mathbf{D}_0)$ and $(\kappa, \beta)' \sim N_2(\mathbf{b}_0, \mathbf{B}_0)$ for their natural conjugacy properties. While previous approaches assumed i.i.d. error terms: $\boldsymbol{\epsilon}_i \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, this approach allows each term to have varying parameters: $\boldsymbol{\epsilon}_i \sim N_2(\mathbf{0}, \boldsymbol{\Sigma}_i)$. Additionally, it assumes that each of the $\boldsymbol{\Sigma}_i$ are i.i.d. from a discrete random distribution G, which is modeled as $G \sim DP(\alpha, G_0)$. Furthermore, G can be integrated out to result in a continuous marginal distribution for $\boldsymbol{\Sigma}_i$. This is also known as a mixture of Dirichlet Processes (MDP).

Conley et al. also consider various established methods for estimation of the error term. Their research found that under departures from normality, the semi-parametric Bayes estimators have smaller root mean square errors (RMSE) than standard classical estimators. Furthermore, the non-parametric Bayes method has identical RMSE for normal errors, and much smaller RMSE for log-normal errors.

Using MCMC sampling techniques, they found that their method produces smaller credibility regions than the classical procedures, under both weak and strong instruments, and especially in the case of non-Normal errors. If the errors are non-Normal, their method may provide efficiency gains; if the errors are Normal, their method definitely does provide efficiency gains. Their results indicate that their methodology is better than the standard Bayesian and classical methods.

3.3 Cholesky Prior

Lopes & Polson (2014) proposed the use of a Cholesky-based prior for $\mathbf{\Sigma}$. Cholesky-based priors have been successfully used for other applications such as longitudinal models (Pourahmadi, 1999), as well as high dimensional stochastic volatility modeling (Lopes, McCulloch, and Tsay, 2011). In this context, this approach allows researchers to model the individual components of the recursive conditional regressions resulting from the Cholesky decomposition of $\mathbf{\Sigma}$. In other words, it addresses the problem seen in using an IW prior for the covariance matrix – the inability to specify varying levels of confidence for the individual components of covariance matrix.

This method utilizes the Cholesky decomposition of $\mathbf{\Sigma}$, which can be represented as follows:

$$\mathbf{\Sigma} = \mathbf{AHA}'$$

In this equation, $\mathbf{A}$ is lower triangular with ones in the main diagonal and lower triangular element $a_{21} = \frac{\sigma_{12}}{\sigma_{11}}$, and $\mathbf{H} = diag(\sigma_{11}, \sigma_{2|1})$. Equivalently, $\sigma_{12} = a_{21}\sigma_{11}$ and $\sigma_{22} = \sigma_{2|1} + \frac{\sigma_{12}^2}{\sigma_{11}}$. Using the common assumption that $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \epsilon_{2i})'$ are i.i.d. $N_2(\mathbf{0}, \mathbf{\Sigma})$, then the distribution of the transformed error terms is as follows:

$$\mathbf{A}^{-1}\boldsymbol{\epsilon}_i \sim N_2(\mathbf{0}, \mathbf{H})$$

and $\boldsymbol{\epsilon}_i \sim N_2(\mathbf{0}, \mathbf{\Sigma})$ can be rewritten using the following conditional regressions (Lopes and Polson, 2014):

$$\begin{aligned} \epsilon_{1i} \sim \quad & N(0, \sigma_{11}) \\ \epsilon_{2i}|\epsilon_{1i} \sim \quad & N(a_{21}\epsilon_{1i}, \sigma_{2|1}) \end{aligned}$$

In these equations, a_{21} represents the strength of correlation between ϵ_{1i} and ϵ_{2i}, and $\sigma_{2|1}$ represents the conditional residual variance.

The prior specifications for the parameters of this approach are similar to that seen for the IW prior in Section 3.1. In fact, the same prior distributions are used for the parameters κ, β, and $\boldsymbol{\delta}$. The difference occurs in its prior specification for the covariance parameters a_{21}, σ_{11}, and $\sigma_{2|1}$. These parameters are assigned the following prior distributions:

$$\begin{aligned} a_{21} \sim \quad & N(\mu_a, \sigma_a) \\ \sigma_{11} \sim \quad & IG(a_{11}, \beta_{11}) \\ \sigma_{2|1} \sim \quad & IG(a_{2|1}, \beta_{2|1}) \end{aligned}$$

Here, μ_a, σ_a, a_{11}, β_{11}, $a_{2|1}$, and $\beta_{2|1}$ are assumed to be specified by the user. By combining these priors with the profile likelihood of $\mathbf{\Sigma}$ in equation (4), a conditional posterior distribution can be calculated. This allows for usage of the same MCMC sampling steps seen in Section 3.1, with only a substitution of the conditional posterior distribution for the error terms.

4. Proposed Prior: Multivariate Normal

In a multivariate Normal setting, the vectorization of the matrix logarithm of the covariance matrix can be shown to have an approximate likelihood with a multivariate Normal form, with respect to the elements of the transformed matrix. A multivariate Normal prior is then happily the conjugate prior for such a likelihood.

Such a prior allows for a flexible specification for the elements of the location parameter (and equivalently, $\mathbf{\Sigma}$). However, the structure of the exact posterior distribution precludes the possibility of an analytical/numerical approach, therefore necessitating the use of

approximation techniques. In particular MCMC sampling can be used with a Metropolis-Hastings accept reject algorithm. The details are discussed below.

4.1 Likelihood Function

Suppose we have a random sample of size n from the structural IV model in equation (3):

$$\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \epsilon_{2i})' \sim N_2(0, \boldsymbol{\Sigma})$. Recall that the joint likelihood of $\mathbf{A}$ and $\boldsymbol{\Sigma}$ is as seen in equation (4):

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\Sigma}|\mathbf{Y}) = (2\pi)^{-n}|\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}tr(\mathbf{W}\boldsymbol{\Sigma}^{-1})\right\}$$

4.1.1 Exact Likelihood

We define the maximum likelihood estimate of $\boldsymbol{\Sigma}$ to be $\mathbf{S}$, such that $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{A}\mathbf{X}_i)(\mathbf{Y}_i - \mathbf{A}\mathbf{X}_i)' = \frac{1}{n}\mathbf{W}$. Then using the joint likelihood for $\boldsymbol{\Sigma}$ and $\mathbf{A}$ in equation (4), the exact profile likelihood for $\boldsymbol{\Sigma}$ is seen to be:

$$\ell(\boldsymbol{\Sigma}|\mathbf{Y}) = (2\pi)^{-n}|\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}tr(\mathbf{S}\boldsymbol{\Sigma}^{-1})\right\}.$$

Now define $\mathbf{C}$ and $\boldsymbol{\Lambda}$ to be the matrix logarithm of $\boldsymbol{\Sigma}$ and $\mathbf{S}$, respectively.

$$\mathbf{C} = \log(\boldsymbol{\Sigma}) = \mathbf{E}[\log(\mathbf{D})]\mathbf{E}'$$

$$\boldsymbol{\Lambda} = \log(\mathbf{S}) = \mathbf{E}_0[\log(\mathbf{D}_0)]\mathbf{E}_0'$$

where $\mathbf{E}$ and $\mathbf{E}_0$ are orthonormal matrices whose columns are the normalized eigenvectors for $\boldsymbol{\Sigma}$ and $\mathbf{S}$, respectively; $\mathbf{D}$ and $\mathbf{D}_0$ are diagonal matrices containing the normalized eigenvalues associated with $\boldsymbol{\Sigma}$ and $\mathbf{S}$, respectively.

Using the equivalence of $|\boldsymbol{\Sigma}| = \exp\{tr(\mathbf{C})\}$, the exact profile likelihood of $\mathbf{C}$ can be expressed as:

$$\ell(\mathbf{C}|\mathbf{Y}) = (2\pi)^{-n} \exp\left\{-\frac{n}{2}tr[\mathbf{C} + \mathbf{S}\exp(-\mathbf{C})]\right\}. \tag{6}$$

4.1.2 Approximate Likelihood

From Bellman (1970), it is known that $\exp(-\mathbf{C})$ can be expressed as a linear Volterra integral equation. Leonard and Hsu (1992, 1999) demonstrated that it can then be approximated by Bellman's iterative solution to the linear Volterra integral equation. This allows for the following approximation to the likelihood function for $\boldsymbol{\gamma}$:

$$\ell^{\star}(\boldsymbol{\gamma}|\mathbf{Y}) \propto |\mathbf{S}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\lambda})'\mathbf{Q}(\boldsymbol{\gamma} - \boldsymbol{\lambda})\right\} \tag{7}$$

where $\boldsymbol{\gamma} = Vec^{*}(\mathbf{C})$ and $\boldsymbol{\lambda} = Vec^{\star}(\boldsymbol{\Lambda})$, where the function Vec^{*} is defined to be the vector of the upper triangular elements of a matrix taken along the diagonal and shifting to the right, and $\mathbf{Q}$ is a (3×3) symmetric almost surely positive definite matrix known as the likelihood information matrix of $\boldsymbol{\gamma}$. Recall that $\mathbf{Q}$ is a function of the normalized eigenvalues and eigenvectors of $\mathbf{S}$

$$\mathbf{Q} = \frac{n}{2}\sum_{i=1}^{2}\mathbf{f}_{ii}\mathbf{f}_{ii}' + n\sum_{i<j}^{2}\sum^{2}\xi_{ij}\mathbf{f}_{ij}\mathbf{f}_{ij}'.$$

and

$$\xi_{ij} = \frac{(d_i - d_j)^2}{d_i d_j [\log(d_i) - \log(d_j)]^2}.$$

and $\mathbf{f}_{ij} = \mathbf{e}_i * \mathbf{e}_j$ represents the (3×1) vector that satisfies the condition $\boldsymbol{\gamma}'(\mathbf{e}_i * \mathbf{e}_j) = \mathbf{e}_i' \mathbf{C} \mathbf{e}_j$. Additionally, d_j is the j^{th} normalized eigenvalue of $\mathbf{S}$ for $j = 1, 2$.

4.2 Prior Structures

The approximate profile likelihood function seen in equation (7) has a multivariate Normal form with respect to $\boldsymbol{\gamma}$, which indicates that the multivariate Normal distribution is the conjugate prior. Suppose our multivariate Normal prior takes the form of

$$\boldsymbol{\gamma} | \boldsymbol{\eta}, \boldsymbol{\Upsilon} \sim N_3(\boldsymbol{\eta}, \boldsymbol{\Upsilon}).$$

Recall that $\boldsymbol{\eta}$ is a (3×1) prior mean location hyperparameter and $\boldsymbol{\Upsilon}$ is a (3×3) prior covariance hyperparameter matrix. This prior specification allows users to specify confidence levels for each element of $\boldsymbol{\gamma}$ in addition to the amount of interdependence between each pair of elements of $\boldsymbol{\gamma}$. If a researcher chose to specify a subset, this can be accomplished by modeling $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\mu})$ and $\boldsymbol{\Upsilon} = \boldsymbol{\Upsilon}(\boldsymbol{\theta})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ are of smaller order than $\boldsymbol{\eta}$ and $\boldsymbol{\Upsilon}$, respectively.

Now if we believe that there are the following two distinct groups in $\boldsymbol{\gamma}$: the variance elements which follow one structure and the covariance elements which follow another structure, we can express this belief as follows:

$$\boldsymbol{\gamma} | \boldsymbol{\mu}, \boldsymbol{\Delta} \sim N_3(\mathbf{J}\boldsymbol{\mu}, \boldsymbol{\Delta}).$$

Thus the prior distributional form is:

$$\pi(\boldsymbol{\gamma} | \boldsymbol{\mu}, \boldsymbol{\Delta}) \propto |\boldsymbol{\Delta}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\gamma} - \mathbf{J}\boldsymbol{\mu})' \boldsymbol{\Delta}^{-1} (\boldsymbol{\gamma} - \mathbf{J}\boldsymbol{\mu}) \right\} \tag{8}$$

where $\mathbf{J}$ is a (3×2) matrix

$$\mathbf{J} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}'$$

Here, we have $\boldsymbol{\mu}$ as a (2×1) vector,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and $\boldsymbol{\Delta}$ is a (3×3) symmetric positive definite matrix

$$\boldsymbol{\Delta} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix}.$$

In this model, μ_1 and σ_1^2 are the location and variance hyperparameters for the variance elements, and μ_2 and σ_2^2 are the location and variance hyperparameters for the covariance elements.

If we are uncertain as to the true values of $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$, a hierarchical prior structure can be considered. We can assume *a priori* that each variable has some known distribution. In the case where no prior information is known about either variable, they are said to have *a priori* a diffuse distribution. For $\boldsymbol{\mu}$, we specify the following prior,

$$\boldsymbol{\mu} \sim N_2(\boldsymbol{\mu}^*, \Psi^*). \tag{9}$$

Note that we can later assume (which we do) that $\boldsymbol{\mu}$ has a diffuse prior ($\pi(\boldsymbol{\mu}) \propto 1$). For the elements of $\boldsymbol{\Delta}$, we specify the following prior,

$$\frac{\nu_i \lambda_i}{\sigma_i^2} \sim \chi^2_{\nu_i}$$

where ν_i is the degree of freedom parameter and λ_i is the scale parameter (both of which will be specified to be quite small). Then the distribution of σ_i^2 can be expressed as follows:

$$\pi(\sigma_i^2) \propto (\sigma_i^2)^{-\frac{\nu_i}{2}-1} \exp\left(-\frac{\nu_i \lambda_i}{2\sigma_i^2}\right). \tag{10}$$

4.3 Exact Posterior

The exact joint posterior distribution is simply a product of the exact profile likelihood function in (6), the prior distribution for $\boldsymbol{\gamma}$ in (8), and the vague prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$ in (9).

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Delta}|\mathbf{Y}) \propto |\boldsymbol{\Delta}|^{-\frac{1}{2}} \exp\left\{-\frac{n}{2} tr[\mathbf{C} + \mathbf{S}\exp(-\mathbf{C})] - \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{J}\boldsymbol{\mu})'\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma} - \mathbf{J}\boldsymbol{\mu})\right\} \pi(\boldsymbol{\Delta})$$

From the equation above, it is evident that the exact joint posterior is not analytically tractable and would thus require numerical techniques for further analysis.

The exact conditional posterior distribution for $\boldsymbol{\gamma}$ can also be calculated, but first requires the calculation of the prior distribution for $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$. This is accomplished by integrating out $\boldsymbol{\mu}$ in the joint prior distribution, which is the product of the prior distribution for $\boldsymbol{\gamma}$ in equation (8) and the vague prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$ in equation (9). This essentially entails integrating out $\boldsymbol{\mu}$ in equation (8), which gives us:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\Delta}) \propto |\boldsymbol{\Delta}|^{-\frac{1}{2}} |\mathbf{J}'\boldsymbol{\Delta}^{-1}\mathbf{J}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}' G^* \boldsymbol{\gamma}\right\} \cdot \pi(\boldsymbol{\Delta}) \tag{11}$$

where $G^* = [I_3 - \mathbf{J}(\mathbf{J}'\boldsymbol{\Delta}^{-1}\mathbf{J})^{-1}\mathbf{J}'\boldsymbol{\Delta}^{-1}]'\boldsymbol{\Delta}^{-1}[I_3 - \mathbf{J}(\mathbf{J}'\boldsymbol{\Delta}^{-1}\mathbf{J})^{-1}\mathbf{J}'\boldsymbol{\Delta}^{-1}]$. Recall that, I_3 is a (3×3) identity matrix, and that this joint prior (with respect to $\boldsymbol{\gamma}$) has a multivariate Normal form.

The exact joint posterior distribution for $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$ is the product of the exact profile likelihood function in equation (6)) and the previously calculated joint prior distribution for $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$ in equation (11).

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\Delta}|\mathbf{Y}) \propto |\boldsymbol{\Delta}|^{-\frac{1}{2}} |\mathbf{J}'\boldsymbol{\Delta}^{-1}\mathbf{J}|^{-\frac{1}{2}} \exp\left\{-\frac{n}{2} tr[\mathbf{C} + \mathbf{S}\exp(-\mathbf{C})] - \frac{1}{2}\boldsymbol{\gamma}' G^* \boldsymbol{\gamma}\right\} \cdot \pi(\boldsymbol{\Delta})$$

From the equation above, it can be seen that the exact conditional posterior distribution is proportional to the exact joint posterior distribution [for $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$] such that

$$\pi(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\Delta}) \propto \pi(\boldsymbol{\gamma}, \boldsymbol{\Delta}|\mathbf{Y}) \propto \exp\left\{-\frac{n}{2} tr[\mathbf{C} + \mathbf{S}\exp(-\mathbf{C})] - \frac{1}{2}\boldsymbol{\gamma}' G^* \boldsymbol{\gamma}\right\} \tag{12}$$

This expression is also not analytically tractable with respect to $\boldsymbol{\gamma}$, and will also require numerical techniques for further analysis.

To be able to use the MCMC technique, we also need to calculate the exact posterior distribution for $\boldsymbol{\Delta}$ conditional on $\boldsymbol{\gamma}$. This posterior distribution is proportional to the product of the exact profile likelihood function in equation (6), the conditional prior distribution in equation (11) for $\boldsymbol{\gamma}|\boldsymbol{\Delta}$, and the prior distribution of $\boldsymbol{\Delta}$ in equation (10). Since the exact profile likelihood function does not depend on $\boldsymbol{\Delta}$, the exact posterior distribution for $\boldsymbol{\Delta}$

conditional on $\boldsymbol{\gamma}$ is simply proportional to the product of the conditional prior distribution in equation (11)) for $\boldsymbol{\gamma}|\boldsymbol{\Delta}$ and the prior distribution for $\boldsymbol{\Delta}$ in equation (10).

$$\begin{aligned}\pi(\boldsymbol{\Delta}|\mathbf{Y},\boldsymbol{\gamma}) \propto \quad & |\boldsymbol{\Delta}|^{-\frac{1}{2}}|\mathbf{J}'\boldsymbol{\Delta}^{-1}\mathbf{J}|^{-\frac{1}{2}}\exp(\boldsymbol{\gamma}'G^{\star}\boldsymbol{\gamma})\cdot\pi(\boldsymbol{\Delta}) \\ \propto \quad & (\sigma_1^2)^{-\frac{r+\nu_1-1}{2}}\exp\left\{-\frac{1}{2\sigma_1^2}\left[\frac{\nu_1\lambda_1}{2\sigma_1^2}+\sum_{i=1}^{2}(\boldsymbol{\gamma}_i-\bar{\boldsymbol{\gamma}}_v)^2\right]\right\} \\ \times \quad & (\sigma_2^2)^{-\frac{q-r+\nu_2-1}{2}}\exp\left\{-\frac{1}{2\sigma_2^2}\left[\frac{\nu_2\lambda_2}{2\sigma_2^2}2\right]\right\}\end{aligned}$$

The exact posterior distribution for $\boldsymbol{\Delta}$ conditional on $\boldsymbol{\gamma}$ is simply the product of two Inverse Gamma distributions:

$$\begin{aligned}\sigma_1^2|\mathbf{Y},\boldsymbol{\gamma} \quad &\sim \quad IG\left(\frac{\nu_1+1}{2},\nu_1\lambda_1+\sum_{i=1}^{2}(\boldsymbol{\gamma}_i-\bar{\boldsymbol{\gamma}}_v)^2\right) \\ \sigma_2^2|\mathbf{Y},\boldsymbol{\gamma} \quad &\sim \quad IG\left(\frac{\nu_2}{2},\nu_2\lambda_2\right)\end{aligned}$$

where $\bar{\boldsymbol{\gamma}}_v = \frac{1}{2}(\gamma_{1^{\cdot}}+\gamma_2)$ is the arithmetic mean of the variance components of $\boldsymbol{\gamma}$. From this equation, we can deduce that the posterior distributions for σ_1^2, σ_2^2 are independent Inverse Gamma density functions. We can set $\nu_1 = \nu_2 = \lambda_1 = \lambda_2 = 1$ so that $\boldsymbol{\delta}$ has a fairly diffuse prior. Then,

$$\begin{aligned}\sigma_1^2|\mathbf{Y},\boldsymbol{\gamma} \sim \quad & IG\left(1,1+\sum_{i=1}^{2}(\boldsymbol{\gamma}_i-\bar{\boldsymbol{\gamma}}_v)^2\right) \\ \sigma_2^2|\mathbf{Y},\boldsymbol{\gamma} \sim \quad & IG\left(\frac{1}{2},1\right).\end{aligned}$$

4.4 Approximate Posterior

The exact conditional posterior distribution is not analytically tractable, so it cannot be used directly to obtain draws for posterior inference. The calculation of an approximate conditional posterior with a convenient form would allow the use of MCMC simulation techniques to simulate "draws" from the exact conditional posterior. An alternative would be to use of the approximate conditional posterior distribution if it has a convenient form.

The approximate joint posterior distribution for $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$ can be calculated by taking the product of the approximate profile likelihood function in (7) and their joint prior distribution in (11),

$$\pi^*(\boldsymbol{\gamma},\boldsymbol{\Delta}|\mathbf{Y}) \propto |\boldsymbol{\Delta}|^{-\frac{1}{2}}|\mathbf{J}'\boldsymbol{\Delta}^{-1}\mathbf{J}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}[(\boldsymbol{\gamma}-\boldsymbol{\lambda})'\mathbf{Q}(\boldsymbol{\gamma}-\boldsymbol{\lambda})+\boldsymbol{\gamma}'\mathbf{G}^*\boldsymbol{\gamma}]\right\}$$

This expression has a form that is similar to that of the multivariate Normal distribution, with respect to $\boldsymbol{\gamma}$. The exponent is re-written with a proportionality taken with respect to the terms that involve $\boldsymbol{\gamma}$ to give the approximate posterior distribution for $\boldsymbol{\gamma}$ conditional on $\boldsymbol{\Delta}$,

$$\pi^*(\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\Delta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)'(\mathbf{Q}+\mathbf{G}^*)(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\} \tag{13}$$

where $\boldsymbol{\gamma}^* = (\mathbf{Q}+\mathbf{G}^*)^{-1}\mathbf{Q}\boldsymbol{\lambda}$. Now, it can be seen that the approximate conditional posterior distribution of $\boldsymbol{\gamma}$ has a convenient form: $\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\Delta} \overset{app}{\sim} N_3(\boldsymbol{\gamma}^*,[\mathbf{Q}+\mathbf{G}^*]^{-1})$.

4.5 Methodology

Consider the IV regression model defined in equations (1) and (2) or equivalently equation (3), the priors for $\boldsymbol{\delta}$, κ and β described in Section 3.1 and the proposed hierarchical prior for $\boldsymbol{\Sigma}$ described in Section 4,2. The complete steps for our MCMC sampling are as follows:

1. Simulate $\sigma_1^{2(t+1)}$, $\sigma_2^{2(t+1)}$ from the conditional posteriors:
   $\sigma_1^2|\mathbf{Y}, \boldsymbol{\gamma}^{(t)} \sim IG\left(1, 1 + \sum_{i=1}^{2}(\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}}_v)^2\right)$
   $\sigma_2^2|\mathbf{Y}, \boldsymbol{\gamma}^{(t)} \sim IG\left(\frac{1}{2}, 1\right)$

2. Simulate $\boldsymbol{\gamma}$: draw a candidate value $\tilde{\boldsymbol{\gamma}}$ from the approximate conditional posterior:

$$\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\Delta}^{(t)} \sim N_3(\boldsymbol{\gamma}^\star, [\mathbf{Q} + \mathbf{G}^*]^{-1})$$

3. Metropolis-Hastings accept reject algorithm. Let:

$$\boldsymbol{\gamma}^{(t+1)} = \begin{cases} \tilde{\boldsymbol{\gamma}} & \text{with probability min}(\rho^\star,1) \\ \boldsymbol{\gamma}^{(t)} & \text{otherwise} \end{cases}$$

   where $\rho^\star = \frac{\pi(\tilde{\boldsymbol{\gamma}}|\mathbf{Y}, \boldsymbol{\Delta}^{(t)})}{\pi^\star(\tilde{\boldsymbol{\gamma}}|\mathbf{Y}, \boldsymbol{\Delta}^{(t)})} \Big/ \frac{\pi(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\Delta}^{(t)})}{\pi^\star(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\Delta}^{(t)})}$ and $\pi(\cdot|\cdot)$ and $\pi^\star(\cdot|\cdot)$ are as defined in equation (12) and (13), respectively

4. Transform $\boldsymbol{\gamma}$ into corresponding $\boldsymbol{\Sigma}$

   (a) If the candidate value $\tilde{\boldsymbol{\gamma}}$ was accepted, transform it into corresponding $\boldsymbol{\Sigma}$.

   (b) If the candidate value $\tilde{\boldsymbol{\gamma}}$ was rejected, keep the current value of $\boldsymbol{\Sigma}$.

5. Simulate (κ, β). The regression parameters (κ, β) have a joint posterior

$$\kappa, \beta|\mathbf{Y}, \boldsymbol{\delta}, \boldsymbol{\Sigma} \sim N_2(b_1, B_1)$$

   where $\mathbf{b}_1$ and $\mathbf{B}_1$ are defined in Section 3.1.

6. Simulate $\boldsymbol{\delta}$. The regression parameter $\boldsymbol{\delta}$ has the posterior

$$\boldsymbol{\delta}|\mathbf{Y}, \kappa, \beta, \boldsymbol{\Sigma} \sim N_2(d_1, D_1)$$

   where $\mathbf{d}_1$ and $\mathbf{D}_1$ are also defined in Section 3.1.

7. Combine κ, β, and $\boldsymbol{\delta}$ into the matrix $\mathbf{A}^{(t+1)}$.

This MCMC scheme is not only easy to implement, but also fairly accurate. It allows researchers to collect samples from a distribution that is difficult to sample from. The main downside to MCMC sampling is the length of time it often takes for sufficient convergence. Current increases in computational power – and subsequently computational speed – have been able to ameliorate this problem to a degree.

5. Conclusion

Instrumental variables regression is an important topic in statistics and, in particular, econometrics. One classical approach to the IV regression problem, 2SLS, is very popular for its efficiency and ease of use. Yet, like all classical approaches, it does not allow the user to include any outside information. Bayesian analysis, on the other hand, allows for prior views about the parameters. It has been shown (Conley et al. 2008) that Bayesian interval estimators perform well against frequentist estimators, even under frequentist performance criteria.

The standard Bayesian approach to modeling the covariance matrix of the error term, seen in equation (5), is with the use of an Inverse Wishart prior. While this allows for the nice properties associated with conjugate priors, it lacks the ability to use all potential prior information. In particular, it only allows for one level of confidence for all of the elements of the covariance matrix. We proposed an alternative that allows for a flexible covariance specification. This approach allows for the specification of varying confidence levels for each element in the variance matrix.

REFERENCES

Anderson, T. W. and Rubin, H. (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63.

Basmann, R. L. (1957), "A Generalized Classical method of Linear Estimation of Coefficients in a Structural Equations," *Econometrica*, 77–83.

Bellman, R. (1970), *Introduction to Matrix Analysis*, New York: McGraw-Hill.

Chao, J. C. and Phillips, P. C. B. (1998), "Posterior Distributions in Limited Information Analysis of the Simultaenous Equations model Using the Jeffreys Prior," *Journal of Econoментrics*, 87, 49–86.

Conley, C. B., Hasen, C. B., McCulloch, R. E., and Rossi, P. E. (2008), "A Semi-parametric Bayesian Approach to thge Instrumental Variable Problem," *Journal of Econometrics*, 144, 276–305.

Drèze, J. H. (1976), "Bayesian Limited Informatiobn Analysis of the Simultaneous Equations Model," *Econometrica*, 1045–1075.

Drèze, J. H. and Morales, J. A. (1976), "Bayesian Full Information Analysis of Aimutaneous Equations," *Journal of the American Statistical Association*, 71, 919–923.

Drèze, J. H. and Richard, J. F. (1983), "Bayesian Analysis of Simutaneous Equation Systems," in *Handbook of Econometrics*, eds. Z. Griliches and M. D. Intrilligator, Ch. 9.

Kleibergen, F. (1997), "Bayesian Simultaneous Equations Analysis Using Equality Restricted Random Variables," *Proceeding of the Section on Bayesian Statistical Science*, 141, 147.

Kleibergen, F. and Zivot, E. (2003), "Bayesian and Classical Approaches to Instrumental Variable Regression," *Journal of Econometrics*, 114, 29–72.

Koopmans, T. C. and Hood, W. C. (1953), "The Estimation of Simultaneous Linear Economic Relations," *Studies in Econometric Method*, 14, 112–119.

Leonard, T. and Hsu, J. S. J. (1992), "Bayesian Inference for a Covariance Matrix," *Annals of Statistics*, 20, 1969–1996.

Leonard, T. and Hsu, J. S. J. (1999), *Bayesian Methods*, Cambridge: Cambridge University Press.

Lopes, H. F., McCulloch, R. E., and Tsay, R. S. (2001), "Cholesky Stochastic Volatility," Technical Report, The University of Chicago Booth School of Business.

Lopes, H. F. and Polson, N. G. (2014), "Bayesian Instrumental Variables: Priors and Likelihoods," *Econometric Reviews*, 33, 100–121.

Pourahmadi, M. (1999), "Joint Mean-covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation," *Biometrika*, 86, 667-690.

Rossi, P. E., Allenby, G. M., and McCulloch, R. E. (2005), *Bayesian Statistics and Marketing*, New York: Wiley.

Theil, H. (1953), "Repeated Least Squares Applied to Complete Equation Systems", The Hague: Central Planning Bureau.

Tsurumi, H. (1990), "Comparing Bayesian and Non-Bayesian Limited Information Estimators, Bayesian and likelihood Methods in Statistics and Econometrics," in *Essays in Honor of George Barnard*, Amsterdam: North-Holland, 179–207.

Zellner, A. (1971), *An Intrroduction to Bayesian Analysis in Econometrics*, New York: Wiley.