GRIPS Discussion Paper 10-32

Bayesian Model Averaging in the Instrumental Variable Regression Model.

By

Gary Koop

Roberto Leon-Gonzalez

Rodney Strachan

Jan 2011

National Graduate Institute for Policy Studies
7-22-1 Roppongi, Minato-ku,
Tokyo, Japan 106-8677

Bayesian Model Averaging in the Instrumental Variable Regression Model*

Gary Koop
University of Strathclyde

Roberto Leon-Gonzalez
National Graduate Institute for Policy Studies

Rodney Strachan
The Australian National University

January 2011

*All authors are fellows of the Rimini Centre for Economic Analysis. The authors would like to thank Frank Kleibergen and other partipants at the European Seminar on Bayesian Econometrics for helpful comments as well as the Leverhulme Trust for financial support under Grant F/00 273/J. Leon-Gonzalez also thanks the Japan Society for the Promotion of Science for financial support (start-up Grant #20830025). Corresponding author: Roberto Leon-Gonzalez, rlg@grips.ac.jp.

ABSTRACT

This paper considers the instrumental variable regression model when there is uncertainty about the set of instruments, exogeneity restrictions, the validity of identifying restrictions and the set of exogenous regressors. This uncertainty can result in a huge number of models. To avoid statistical problems associated with standard model selection procedures, we develop a reversible jump Markov chain Monte Carlo algorithm that allows us to do Bayesian model averaging. The algorithm is very flexible and can be easily adapted to analyze any of the different priors that have been proposed in the Bayesian instrumental variables literature. We show how to calculate the probability of any relevant restriction (e.g. the posterior probability that over-identifying restrictions hold) and discuss diagnostic checking using the posterior distribution of discrepancy vectors. We illustrate our methods in a returns-to-schooling application.

Keywords: Bayesian, endogeneity, simultaneous equations, reversible jump Markov chain Monte Carlo.

JEL Classification: C11, C30

1 Introduction

For the regression model where all potential regressors are exogenous, a large literature[1] has arisen to address the problems caused by a huge model space. That is, the number of models under consideration is typically 2^K where K is the number of potential regressors. With such a huge model space, there are many problems with conventional model selection procedures (e.g. sequential hypothesis testing procedures run into pre-test problems). Bayesian model averaging (BMA) can be used to avoid some of these problems. However, the size of the model space means that carrying out BMA by estimating every model is typically computationally infeasible. Accordingly, an algorithm which simulates from the model space (e.g. the Markov chain Monte Carlo model composition algorithm of Madigan and York, 1995) must be used. In the case of the regression model with exogenous regressors, such methods are well-developed, well-understood and are increasingly making their way into empirical work. However, to our knowledge, there are no comparable papers for the empirically important case where regressors are potentially endogenous and, thus, instrumental variable (IV) methods are required.[2] The purpose of the present paper is to fill this gap.

Inference about structural parameters in the IV regression model requires the formulation of assumptions whose validity is often uncertain. A useful representation of the model is the incomplete simultaneous equations model (see, for example, Hausman, 1983). Within this representation, the most crucial assumptions relate to the set of instruments and the rank condition for identification (Greene, 2003, p. 392). In addition to these, one has to decide how many regressors to include, and which of these are potentially endogenous. This can lead to a huge model space and, thus, similar issues arise as for the regression model with exogenous regressors. In practice, researchers typically try different specifications until a set of restrictions (i.e. a particular choice of instruments, exogenous and endogenous regressors) passes a battery of misspecification tests (e.g. Anderson and Rubin, 1949, 1950, Hausman,

[1] See, among many others, Fernandez, Ley and Steel, 2001 and the references cited therein.

[2] Two related papers are Cohen-Cole, Durlauf, Fagan, and Nagin (2009) and Eicher, Lenkoski and Raftery (2009) but the model space in these papers is small and, hence, simulation methods from the model space are not required. Furthermore, the approach of these papers (averaging of two-staged least squares estimates using BIC-based weights) does not have a formal Bayesian justification.

1983, Sargan, 1958). Given the large number of possible models, the repeated application of diagnostic tests will result in similar distorted size and power properties as arise in the regression model with exogenous regressors. Since estimates of structural estimates that rely on incorrect identification restrictions can result in large biases, the consequences of these problems can be substantive. BMA can be used to mitigate such problems. But the size of the model space often precludes estimation of all models. This leads to a need for computational methods which simulate from the model space. A contribution of the present paper is to design a reversible jump Markov chain Monte Carlo algorithm (RJMCMC, see Green, 1995 or Waagepetersen and Sorensen, 2001) that explores the joint posterior distribution of parameters and models and thus allows us to do BMA. This allows us to carry out inference on the structural parameters that, conditional on identification holding, accounts for model uncertainty. Furthermore, our algorithm allows for immediate calculation of the posterior probability associated with any restriction, model or set of models. Thus, we can easily check the validity of identifying restrictions (or exogeneity restrictions, etc.) by calculating the posterior probability of these restrictions. Alternatively, we can use the BMA posterior distribution of discrepancy vectors and functions (Zellner, Bauwens and van Dijk, 1988) in order to shed light on the validity of instruments.

In our applications, we find that standard versions of RJMCMC algorithms (e.g. adapting the RJMCMC methods for seemingly related regression, SUR, models developed by Holmes, Denison and Mallick, 2002, to the IV case) can perform poorly, remaining stuck for long periods in models with low posterior probability. To improve the performance of our RJMCMC algorithms, we borrow an idea from the simulated tempering literature and augment our model space with so-called cold models. The cold models are similar to the models of interest (called hot models) but are simplified in such a way that the RJMCMC algorithm makes very rapid transitions between cold models. As suggested by the simulated tempering literature, we find that this strategy helps the algorithm escape from local modes in the posterior.

The RJMCMC algorithm we propose is very flexible and can be easily adapted to handle any of the popular approaches to Bayesian inference in IV models. To illustrate this, we describe in detail how the algorithm works in the context of three popular Bayesian approaches to instrumental variables and reduced rank regression. These are the classic approach of Drèze (1976) as well as the modern approaches of Kleibergen and van Dijk (1998) and

Strachan and Inder (2004)[3]. We also show how, if desired, the RJMCMC algorithm can be easily coded to produce results for all three (or more) priors by running the algorithm just once.

Section 2 describes the model space we consider. Section 3 describes the algorithm with complete details being included in a Technical Appendix. Section 4 explains how to obtain the BMA posterior distribution of discrepancy measures for model diagnostics proposed by Zellner, Bauwens and van Dijk (1988). Section 5 applies our methods to a returns-to-schooling example based on Card (1995) and Section 6 concludes.

2 Modelling Choices in the Incomplete Simultaneous Equations Model

We will work with the incomplete simultaneous equations model, which takes the form:

$$\begin{aligned} y_{1i} &= \gamma' y_{2i} + \beta' x_i + u_{1i} \\ y_{2i} &= \Pi_{2x} x_i + \Pi_{2z} z_i + v_{2i} \end{aligned} \tag{1}$$

where $y_{1i} : 1 \times 1$, $y_{2i} : m \times 1$, $x_i : k_{1j} \times 1$, $z_i : k_{2j} \times 1$, $i = 1, ..., N$. The errors are normal with zero means and are uncorrelated over i. We assume

$$E\left(x_i \begin{pmatrix} u_{1i} \\ v_{2i} \end{pmatrix}'\right) = 0 \text{ and } E\left(z_i \begin{pmatrix} u_{1i} \\ v_{2i} \end{pmatrix}'\right) = 0.$$

The reduced form version of this model can be written as:

$$y_i = \Pi_x x_i + \Pi_z z_i + v_i \tag{2}$$

[3]We use a proper prior version of the improper prior used by Drèze (1976), as in the subsequent papers of Drèze and Richard (1983) and Zellner, Bauwens and van Dijk (1988). With respect to the prior by Strachan and Inder (2004), we will use a parameter-augmented version of it similar to that used by Koop, Leon-Gonzalez and Strachan (2010).

where $y_i = (y_{1i}, y'_{2i})'$, $v_i = (v_{1i}, v'_{2i})'$ and:

$$\begin{aligned}
\Pi_x &= \begin{pmatrix} \pi_{1x} \\ \Pi_{2x} \end{pmatrix} = \begin{pmatrix} \gamma'\Pi_{2x} + \beta' \\ \Pi_{2x} \end{pmatrix}, \qquad \Pi_z = \begin{pmatrix} \pi_{1z} \\ \Pi_{2z} \end{pmatrix} = \begin{pmatrix} \gamma' \\ I_m \end{pmatrix} \Pi_{2z} \\
\Omega &= E(v_i v_i') \qquad \Sigma = E\left(\begin{pmatrix} u_{1i} \\ v_{2i} \end{pmatrix} \begin{pmatrix} u_{1i} & v'_{2i} \end{pmatrix} \right) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{bmatrix} \\
\Omega &= \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \Omega_{22} \end{bmatrix} = \begin{pmatrix} 1 & \gamma' \\ 0 & I_m \end{pmatrix} \Sigma \begin{pmatrix} 1 & 0 \\ \gamma & I_m \end{pmatrix} \\
\Pi_x &: (m+1) \times k_{1j} \qquad\qquad \Pi_z : (m+1) \times k_{2j}
\end{aligned}$$

The subindex j stands for the j^{th} model, and j varies from 1 to N^{mod}, where N^{mod} is the total number of models. To avoid notational clutter, we will not attach j subindices to parameter matrices although, of course, these will vary over models.

When using this model, there are many sources of uncertainty over identification that arise. Assuming $\sigma_{12} \neq 0$, we can solve for the parameters (β', γ') from the reduced form matrix

$$\widetilde{\Pi} = [\Pi_x \quad \Pi_z]$$

through the relations

$$\pi_{1x} - \gamma'\Pi_{2x} = \beta' \text{ and} \tag{3}$$

$$\pi_{1z} - \gamma'\Pi_{2z} = 0. \tag{4}$$

If we are able to solve (4) for γ, we can subsequently solve for β using (3). Solving for γ depends upon the rank of the matrix Π_z. If $k_{2j} = m$ and $rank\,(\Pi_z) = m$ then there is a unique solution $\gamma' = \pi_{1z}\Pi_{2z}^{-1}$ and the equation is just identified. If $k_{2j} > m$ and $rank\,(\Pi_z) = m$ then there are many solutions such as $\gamma' = \pi_{1z}\Pi_{2z}^{*\prime}\left(\Pi_{2z}^{*}\Pi_{2z}^{*\prime}\right)^{-1}$ where Π_{2z}^{*} is constructed from any set of $k^* \geq m$ linearly independent columns of Π_{2z}. In this case, the equation is over-identified. If $k_{2j} < m$ then $rank\,(\Pi_z) < m$, so there are no solutions and the equation is under-identified.

Uncertainty over identification can also result from uncertainty over what variables in y_{2i} are endogenous and what variables in z_i are not valid instruments. If we relax the earlier assumption on σ_{12} to allow for $\sigma_{12} = 0$, which implies y_{2i} is exogenous, then we have additional solutions for γ from $\gamma' = \omega_{12}\Sigma_{22}^{-1}$ and the condition $\sigma_{12} = 0$ needs to be taken into account when

determining whether (β', γ') is just or over-identified. A further complication arises if elements of γ or σ_{12} are zero as these restrictions imply elements of y_{2i} are exogenous. This effectively changes the value of m, increasing the number of identifying restrictions in (4) and, hence, the conditions for under, just and over identification. Note also that, if $k_{2i} > m$ and rows of the $k_{2i} \times m$ matrix Π_{2z} are zero, or, more generally, if $rank\,(\Pi_{2z}) = m_i < m$, then not all elements of z_i may be regarded as valid instruments. In this case, we can then represent Π_{2z} as the product of two lower dimensional matrices, $\Pi_{2z} = \underline{\Pi}_{2z}\varrho$ where $\underline{\Pi}_{2z}$ is $m \times m_i$ and ϱ is $m_i \times k_{2i}$ both full rank. The valid instruments are then ϱz_i.

Furthermore, if elements of β are zero, then this gives us more equations of the type (4) and few equations of the type (3), again affecting the identification status of (β', γ').

In this paper, we consider a model space which includes all the over-identified and just-identified models (see below for a discussion of non-identified models). These are the models in which $k_{2j} \geq m$ and Π_{2z} has full rank. Models in this category differ according to the following aspects:

- Set of instruments: The variables in z_i are a subset of a larger group of potential instruments denoted by Z^*. There is uncertainty as to which subset of Z^* should enter in the model and hence uncertainty about the column dimension of the matrix Π_{2z}.

- Variables in x_i: x_i is a subset of $Z^* \cup X^*$, where X^* is the set of all potential regressors that are not allowed to be instruments. Uncertainty about what variables enter x_i implies uncertainty over the elements of β.

- Restrictions on the coefficients of endogenous regressors: some coefficients in γ might be restricted to be zero.

- Exogeneity: some of the covariances between u_{1i} and v_{2i} might be zero; that is, there is uncertainty about the elements of σ_{12}.

Note that researchers typically have some exogenous variables that they are certain cannot be instruments (and thus, we introduce X^* as above). However, they are typically interested in checking the validity of all exclusion restrictions (i.e. restrictions that instruments do not enter the structural equation) and, for this reason, our set of potential exogenous regressors in

our equation of interest will include all the potential instruments (i.e. we have $x_i \subseteq Z^* \cup X^*$).

Note that just-identified models are observationally equivalent to (non-identified) full rank models (i.e. models where Π_z has full rank) in which all exclusion restrictions fail. In this sense we are also including non-identified full rank models in our analysis. A problem arises in that different just-identified models will all yield the same full rank model and, thus, are observationally equivalent. That is, full rank models take the form of unrestricted SUR models. But different just-identified models will always have the same unrestricted SUR reduced form (and, thus, yield the same marginal likelihood and be observationally equivalent). Over-identified models will impose restrictions on the coefficients in the reduced form SUR and break this observational equivalence problem. But the observational equivalence of different just-identified models raises the question of how they should be included in a BMA exercise. As an example, consider a reduced form unrestricted SUR model with two equations and two explanatory variables, z_1 and z_2. This reduced form is consistent with a just-identified model where z_1 is the single valid instrument for the first equation. But it is also consistent with a just-identified model where z_2 is the single valid instrument. Should we treat these as two different models weighted equally when doing model averaging? This is a possible strategy that could be done. Or one might prefer to simply treat the two models as one model. Furthermore, as the identifying assumption cannot be tested in the just-identified case, one might decide not to use just-identified models when constructing BMA estimates of structural parameters. But of course just-identified models could be included if desired, and this is what we do in our empirical analysis.

If some elements in γ (and/or Σ) are restricted to be zero then this increases the degree of over-identification such that some models with $k_{2j} \leq m$ may, by these restrictions, become over-identified. However, all of our over-identified models have $k_{2j} > m$. This condition is necessary because a model with some zero restrictions on γ and with fewer than m instruments (even though its parameters are identified) is observationally equivalent to a model in which all elements of γ are different from zero but Π_{2z} has reduced rank. Thus, we consider over-identified models to be those with $k_{2j} > m$, regardless of the restrictions on γ or Σ.

In a subsequent section, we present empirical work based on the classic returns-to-schooling paper of Card (1995) and associated data set. Details are provided in the Data Appendix. However, to make concrete our modelling

framework it is convenient to begin introducing the empirical example here. This cross-sectional data set has 13 potential instruments (this is the set Z^*), 4 endogenous variables (hence $m = 3$), and 27 exogenous regressors (X^*). The structural equation of interest has the log of the wage as the dependent variable (y_{1i}). The key structural parameter of interest is the return to schooling which is an element of γ since years of education is treated as endogenous (i.e. it is an element of y_{2i}).

Consider first over-identified models. Our model space involves[4] C_j^{13} for $j = 3, .., 13$ combinations for each number of instruments. There are 40 potential explanatory variables in $Z^* \cup X^*$, but if a model includes an element of Z^* as an instrument then this element cannot also be in X^*. Hence, we obtain

$$N^A = \sum_{j=3}^{13} 2^{40-j} C_j^{13}$$

over-identified models if we ignore exogeneity restrictions and restrictions on γ. But there are 2^m of each of these resulting in $64N^A$ over-identified models. Adding all these models together yields more than 10^{16} models. This calculation is presented to clarify our class of models and reinforce the point that in common empirical problems it is easy to have a model space which is huge.

3 RJMCMC Algorithms in the Incomplete Simultaneous Equations Model

If the number of models is small (e.g. if the researcher is clear on which variables are potential instruments and their number is small), then conventional methods of Bayesian analysis can be used. That is, the researcher can simply carry out a posterior analysis of every single model. However, in many cases (such as the one used in our empirical work), the number of potential instruments or other modelling choices implies that the model space is huge. In this case, the conventional strategy of carrying out posterior analysis will be computationally infeasible. Such considerations motivate why we wish to

[4] C_c^b denotes "b choose c": the number of sets of c elements chosen without replacement from a set of b elements.

develop an RJMCMC algorithm to sample from the joint posterior defined over the parameter and the model spaces. In this section, we will offer an informal and intuitive explanation of our RJMCMC algorithms with complete details being given in the Technical Appendix.

In this informal section, we will adopt notation where the data is denoted by Y, we have M_j for $j = 1, .., N^{\text{mod}}$ models and each model depends on parameters Π_j which determine the conditional mean of the incomplete simultaneous equations model (i.e. $\Pi_j = (\beta', \gamma', vec(\Pi_{2x})', vec(\Pi_{2z})')'$) and Σ_j is the error covariance matrix. As above, we will suppress the j subscripts and refer to our algorithm as taking draws from the posterior of (Π, Σ, M). We will denote the r^{th} draw from this posterior as $(\Pi^{(r)}, \Sigma^{(r)}, M^{(r)})$ for $r = 1, .., R$. Given draws from this posterior we can do BMA for any posterior feature of interest (e.g. conditional on identification holding, the structural form parameters are a function of Π and we can derive their BMA posterior) or calculate the posterior probability of any subset of the models (e.g. we can calculate the posterior probability associated with over-identified models).

3.1 An RJMCMC Algorithm for the SUR Model

To explain our algorithm, we begin by describing the algorithm of Holmes, Denison and Mallick (2002), hereafter HDM, for doing BMA in the SUR model. If we restricted our model space to over-identified models and adopt the prior of Drèze (1976), we can use this algorithm. However, for reasons explained below, in general this will not result in a good algorithm for IV models. Nevertheless, it is the base on which we build, so we explain this approach here.

HDM motivate their algorithm as an MCMC algorithm providing a sample from $p(\Pi, \Sigma, M|Y)$ by sequentially drawing from:

1. $p(M|Y, \Sigma)$
2. $p(\Pi|Y, \Sigma, M)$
3. $p(\Sigma|Y, \Pi, M)$

HDM assume that, in any model, the prior $p(\Pi, \Sigma) = p(\Pi|\Sigma)\, p(\Sigma)$ is such that $p(\Pi|\Sigma)$ is normal and $p(\Sigma)$ is inverted-Wishart. Under these assumptions, $p(\Sigma|Y, \Pi, M)$ and $p(\Pi|Y, \Sigma, M)$ can be obtained using textbook results for the SUR model (see, e.g., Koop, 2003, pp. 137-142). Thus, steps

2 and 3 in their algorithm are straightforward. Step 1 proceeds by drawing a candidate model M^* and accepting it with probability:

$$\min\left\{\frac{p(Y,\Sigma|M^*)}{p(Y,\Sigma|M^{(r-1)})}\frac{p(M^*)}{p(M^{(r-1)})},1\right\} \tag{5}$$

where:

$$p(Y,\Sigma|M)=\int p(\Pi,\Sigma|M)p(Y|\Pi,\Sigma,M)d\Pi. \tag{6}$$

Note that the densities in the acceptance probability are evaluated at the observed data, Y, and $\Sigma^{(r-1)}$. HDM draw models conditionally on Σ in the SUR model because, while $p(Y|M)$ does not have an analytical form, for HDM's choice of prior, $p(Y,\Sigma|M)$ can be evaluated analytically. This explains why our algorithms also draw models conditional on Σ. As we shall see, it is this inability to analytically integrate Σ out of $p(Y,\Sigma|M)$ which causes problems with the HDM algorithm and motivates our more sophisticated algorithm based on simulated tempering.

The HDM algorithm can also be interpreted as an RJMCMC algorithm which draws from $p\left(\Pi,M|Y,\Sigma\right)$ and $p\left(\Sigma|Y,\Pi,M\right)$. To sample from $p\left(\Pi,M|Y,\Sigma\right)$ an RJMCMC algorithm would proceed by specifying a density for generating candidate models, M^*. In general, this candidate density would take the form $q(M^*|\Sigma,M^{(r-1)})$. Then a candidate draw Π^* would be taken from $q(\Pi^*|\Sigma,M^*)$. An RJMCMC algorithm would then accept the candidate draw $\left(\Pi^{(*)},M^{(*)}\right)$ with an appropriate acceptance probability. If accepted, we have $\left(\Pi^{(r)},M^{(r)}\right)=\left(\Pi^{(*)},M^{(*)}\right)$. If not, then $\left(\Pi^{(r)},M^{(r)}\right)=\left(\Pi^{(r-1)},M^{(r-1)}\right)$.

For the SUR model, it can be shown that choosing $q(\Pi^*|\Sigma,M^*)=p\left(\Pi^*|Y,\Sigma,M^*\right)$ leads to the most efficient RJMCMC algorithm. As we have seen, since HDM use a normal prior for Π, $p\left(\Pi^*|Y,\Sigma,M^*\right)$ has a textbook analytical form. Choosing a type of symmetric random walk for $q(M^*|\Sigma,M^{(r-1)})$, the RJMCMC acceptance probability turns out to be precisely (5). Thus, HDM's algorithm is an RJMCMC algorithm, an interpretation we build on below.

There are two problems with directly using HDM's approach in the incomplete simultaneous equations model. First, the priors used by Bayesians in IV problems rarely involve a normal prior for Π and thus, the analytical results used by HDM are not available. The second problem is more subtle and relates to the fact that the algorithm draws models conditionally on Σ. This problem is worth explaining as it helps to motivate our algorithm.

The problem arises since (5) depends on $p(Y,\Sigma^{(r-1)}|M^*)$ and $p(Y,\Sigma^{(r-1)}|M^{(r-1)})$, but $\Sigma^{(r-1)}$ is drawn conditionally on $M^{(r-1)}$. In practice, this can mean

$p(Y, \Sigma^{(r-1)}|M^*)$ is much lower than $p(Y, \Sigma^{(r-1)}|M^{(r-1)})$ even if M^* is a much better model than $M^{(r-1)}$. Speaking informally, even if $M^{(r-1)}$ is a "bad" model and M^* is a "good" model, $\Sigma^{(r-1)}$ is typically drawn in an area of high posterior probability under $M^{(r-1)}$. So $\Sigma^{(r-1)}$ is "good" for $M^{(r-1)}$ (and, thus, $p(Y, \Sigma^{(r-1)}|M^{(r-1)})$ is large) but may be very "bad" for M^* (and, thus, $p(Y, \Sigma^{(r-1)}|M^*)$ may be low). If enough draws are taken from the algorithm it will eventually escape from such local modes, but in practice we have found it can remain stuck for long periods. Put another way, in the IV case, the model can be highly correlated with Σ and this can lead to very slow convergence.

3.2 An RJMCMC Algorithm for the IV Model of Drèze (1976)

Drèze's (1976) seminal paper on the Bayesian analysis of simultaneous equations models provides the starting point for developing an algorithm for doing BMA in our modelling framework. Drèze (1976) does not consider as extensive a model space as we do, so some extensions of his prior are required (see Technical Appendix for details). But the main element of his approach is the use of a normal prior for $\Pi = (\beta', \gamma', vec(\Pi_{2x})', vec(\Pi_{2z})')$. Thus, the prior setup is the same as in HDM and, thus, in theory the HDM algorithm could be used with the Drèze prior. However, the preceding sub-section showed how the HDM algorithm for SUR models can work poorly.

We stressed the role of Σ in the breakdown of the HDM approach. The strategy we propose to surmount this problem is similar in spirit to the method of simulated tempering (ST) developed by Marinari and Parisi (1992) and Geyer and Thompson (1995). This method was designed to improve the performance of an MCMC algorithm that samples from the posterior distribution of a single model, but we use it in our multiple model case. As in the ST method, we expand the model space with so-called 'cold models'. These cold models are of no intrinsic interest to the researcher, whereas the models that are of interest which we have defined in Section 2 are called 'hot models'. Only the draws from the hot models are included in calculating posterior features of interest (e.g. posterior probabilities for each model, posteriors for structural parameters, etc.). But, if the set of cold models is carefully chosen, their addition can greatly facilitate movement between different hot models. We choose our set of cold models to over-come the

problem noted above, which arises since M and Σ can be so highly correlated.

Complete details are provided in the Technical Appendix. But the key insight is that, if we can find cold models where $p(Y|M)$ can be calculated analytically then, the algorithm will tend to switch easily between cold models since the RJMCMC acceptance probability will no longer depend on $p(Y, \Sigma|M)$ as in (5), but rather on $p(Y|M)$. The problems noted above caused by the conditioning on Σ will be removed. Furthermore, if each cold model is similar to a hot model then the algorithm should switch easily between hot and cold models as well. Our cold models satisfy these requirements.

To be precise, each of our hot models is defined by a likelihood function, a normal prior for Π and an inverted Wishart prior for Σ. Each of our cold models is based on an approximation to the posterior. Formally, we approximate the marginal posterior $p(\Pi_{2z}|Y)$ with a multivariate Student density centered at the maximum likelihood estimate.[5] We combine this with $p(\beta, \gamma, \Pi_{2x}, \Sigma|\Pi_{2z}, Y)$, which is known analytically, to obtain an approximation of the posterior of all unknown parameters and of $p(Y|M)$. See the Technical Appendix for details of our approximation.

As shown below, we have found this algorithm to work well and avoid the problems associated with the algorithm of HDM. There are several minor complications (e.g. treating models with exogeneity restrictions or restrictions on γ) that must be dealt with. Full details of this algorithm, including a treatment of such complications, is provided in the Technical Appendix.

3.3 An RJMCMC Algorithm for the IV Model with Other Priors

In recent years, there have been several alternative priors proposed for the incomplete simultaneous equations model. Two prominent approaches are outlined in Kleibergen and van Dijk (1998) and Strachan and Inder (2004).[6] We will not explain these approaches here (see Technical Appendix for precise formulae), nor motivate their advantages over Drèze (1976). Rather we outline a MCMC strategy for use when we have a prior $p^*(\Pi, \Sigma)$ which is

[5] Note that because Π_{2z} is a reduced form matrix, the asymptotic approximation we use is not affected by the problem of weak instruments.

[6] This latter paper is for the error correction model, but the structure of that model is identical to the incomplete simultaneous equations model.

different from the prior used in Drèze (1976) which we denote by $p^D(\Pi, \Sigma)$. A problem with the use of more general priors is that neither $p(Y|M)$ nor $p(Y, \Sigma|M)$ will be available in closed form. Recall that these are crucial ingredients in our RJMCMC acceptance probabilities. However, it is possible to extend our previous ST algorithm with an extra layer of hot models (let us call these "super-hot models" to distinguish them from our previous hot models which are based on Drèze's prior).

Our algorithm begins with the cold and hot models exactly as in the preceding sub-section. Corresponding to each hot model, we will add a super-hot model which is identical to the hot model, except that it uses $p^*(\Pi, \Sigma)$ instead of $p^D(\Pi, \Sigma)$ as a prior. In other words, the posterior for each super-hot model equals the posterior for a hot model times $\frac{p^*(\Pi,\Sigma)}{p^D(\Pi,\Sigma)}$ and this ratio of priors is the important factor in the acceptance probability. Because of this, in our algorithm, transitions between hot and super-hot models are conditional on both Π and Σ, but in practice we have found this not to be a problem since the hot and super-hot models tend to be very similar to one another.

Note that this algorithm produces draws from cold, hot and super-hot models. In this sense, it is an algorithm that can be used to handle several priors in one RJMCMC run. That is, if we just retain the draws from the super-hot models, then we are doing BMA using one of the alternative priors. If we just retain draws from the hot models, then we are doing BMA using the prior of Drèze (1976). If we just retain the draws from the cold models, then we are doing BMA using an approximation to $p(Y|M)$ and to the posterior density of parameters.

For complete details see the Technical Appendix.

4 Model Comparison and Diagnostics

The posterior probability of any desired restriction can be calculated in a straightforward manner using output from the RJMCMC algorithm. For instance, the posterior probability of over-identification might be of interest to the researcher. This will simply be the proportion of draws taken from over-identified models The posterior probability of each exogeneity restriction or that each element of γ equals zero can be calculated in the same manner. In our empirical work we illustrate how these are done. However, in the Bayesian IV literature, Zellner, Bauwens and van Dijk (1988) have

proposed various discrepancy vectors and functions that measure the extent to which restrictions (e.g. over-identifying restrictions) are in error. In our context, it is natural to consider the BMA posterior of these discrepancy measures. We will consider discrepancy measures for over-identification and under-identification.

Following Zellner, Bauwens and van Dijk (1988) we decompose Π_z as $\Pi_z = (\pi'_{1z}, \Pi'_{2z})'$ such that π_{1z} has only one row. Then we define the Generalized Indirect Least Squares (GILS) of γ as: $\gamma^* = (\Pi_{2z}\Pi'_{2z})^{-1}\Pi_{2z}\pi'_{1z}$, and define $\widetilde{\Delta}_o = (\pi'_{1z} - \Pi'_{2z}\gamma^*)$. The discrepancy function that we use is $\widetilde{d}_o = \left(\widetilde{\Delta}'_o\widetilde{\Delta}_o\right)/k_{2j}$, where k_{2j} is the number of instruments. If the over-identifying restrictions hold, $\widetilde{d}_o$ will be zero. As noted by Zellner, Bauwens and van Dijk (1988) the posterior of $\widetilde{d}_o$ can be obtained by directly drawing from the (matrix Student) posterior distribution of Π_z in the unrestricted full rank model, and calculating $\widetilde{d}_o$ for each value of Π_z that is drawn. In our case, for each over-identified model visited by our RJMCMC algorithm, we will draw one value of $\widetilde{d}_o$ from the (matrix Student) posterior distribution of Π_z in the corresponding unrestricted full rank model. By doing this we can reconstruct the BMA posterior density of $\widetilde{d}_o$. One way to assess whether the BMA posterior of $\widetilde{d}_o$ is close to zero is by comparing it with draws from the prior of $\widetilde{d}_o$. These can be obtained by getting a draw from the prior of Π_z (and then transforming to a draw of $\widetilde{d}_o$) for each of the models visited by the algorithm.[7] A BMA posterior of $\widetilde{d}_o$ that is closer to 0 than the BMA prior signals that the data supports the over-identifying restrictions.

Zellner, Bauwens and van Dijk (1988) did not explicitly provide discrepancy measures for under-identification, but it is possible to adapt their approach to this case. The problem of under-identification arises when the rank of Π_{2z} is less than m, in which case (4) cannot be solved for γ. The $m \times k_{2j}$ matrix Π_{2z} has reduced rank if it can be written as the product of a $m \times (m-1)$ times a $(m-1) \times k_{2j}$ matrix. Using a linear normalization restriction (Johansen, 1995, p. 72), a lower rank Π_{2z} could be written as:

$$\Pi_{2z} = \begin{pmatrix} \delta' \\ I_{m-1} \end{pmatrix} \varrho$$

[7] If the prior of Π_z depends on Ω and the prior of Ω is improper, then we cannot draw from the prior. Instead of this one could draw Ω from the posterior first and then Π_z from the conditional prior of Π_z given Ω.

where $\delta : (m-1)\times 1$, $\varrho : (m-1)\times k_{2j}$. Thus, similarly to the over-identification discrepancy measure, to define the under-identification discrepancy measure, $\widetilde{d}_u$, we decompose the full rank matrix Π_{2z} as $\Pi_{2z} = \left(\pi'_{1,2z}, \varrho'\right)'$ such that $\pi_{1,2z}$ has only one row, and define $\delta^* = \left(\varrho\varrho'\right)^{-1}\varrho\pi'_{1,2z}$, and $\widetilde{\Delta}_u = (\pi'_{1z} - \Pi'_{2z}\delta^*)$. The under-identification discrepancy function that we use is $\widetilde{d}_u = \left(\widetilde{\Delta}'_u\widetilde{\Delta}_u\right)/k_{2j}$. Values of $\widetilde{d}_u$ near 0 will signal that the under-identification restriction does not hold.

5 An Application to Estimating the Returns to Schooling

This empirical illustration is based on Card (1995). Our Data Appendix provides details about the data including definitions of all variables and what type of variable each is (i.e. whether each variable is in y, X^* of Z^*). As noted at the end of Section 2, our model space for this application will include approximately 10^{16} models. In the following we will consider all models equally likely, and the appendix describes the choices for the other prior parameters. As stressed previously, with this many models, it is computationally infeasible to do BMA by carrying out Bayesian inference in every model and then averaging across models using marginal likelihoods. Thus, with the full model space, we cannot compare our RJMCMC to a conventional BMA strategy. Accordingly, before we present a full empirical analysis using all the models, we provide such a comparison using a reduced set of models.

5.1 Comparing RJMCMC to Conventional BMA

In this sub-section, we compare our RJMCMC algorithm to conventional BMA using a reduced set of 7814 models that result from considering instrument uncertainty only. That is, Z^* will consist of 13 regressors that must be allocated to either x_i or z_i, with the restriction that z_i must contain at least 4 of them (which is the minimum for the model to be over-identified). X^* consists of 27 regressors and a constant and these will always enter in x_i. No restrictions on either Σ or γ are considered. By reducing the number of models we are able to calculate the marginal likelihood of each model individually. We can then compare the results to those provided by our RJMCMC algorithm to evaluate the accuracy of our algorithm.

Remember that our algorithm involves cold ($T = 0$), hot ($T = 1$) and super-hot ($T = 2$) models and we refer to these as being of different "temperatures". As discussed previously, draws from the cold and hot models can be used to carry out Bayesian inference under a normal approximation and the prior of Drèze (1976). Here $T = 2$ represents models with the prior of Kleibergen and van Dijk (1998),[8] but (in order to illustrate a variety of approaches) in the empirical application it will represent models with priors in the style of Strachan and Inder (2004).

In order not to unfairly advantage RJMCMC, we deliberately choose a poor starting value for this algorithm: the model we knew had the smallest posterior probability amongst the hot models.[9] In 20000 iterations[10], the number of distinct models visited by our algorithm was 47 and 44 for cold and hot temperatures, respectively. The posterior probability of a model is the proportion of times that the algorithm draws the model.

Tables 1, 2 and 3 present the best models for each temperature. In all cases, these cover about 92% of the total probability mass. We can see that for each temperature the posterior probability given by our algorithm is very close to the one calculated over the whole set of 7814 models. Thus, our RJMCMC algorithm is working well. Although posterior probabilities are quite similar across temperatures, note that there are some differences between cold and hot temperatures, and that our algorithm is able to capture well these differences. For example, the model that excludes only instrument 9 from the actual set of instruments is ranked 2^{nd} when $T = 1$ (with posterior probability 18%) but is ranked 4^{th} when $T = 0$ with posterior probability being only 6% and is not included in the ranking when $T = 2$. On the other hand, the model that uses all 13 variables in Z^* as actual instruments is the best model for all temperatures 0, 1 and 2 (with at least 41% posterior probability).

Table 4 shows the posterior probability that each of the variables in Z^*

[8] Since the MCMC methods for the Kleibergen and van Dijk (1998) approach are computationally demanding, we reduce the set of models that must be estimated. We first run our algorithm with only 2 temperatures ($T = 0$ and $T = 1$) and find the models visited by this algorithm. We evaluate the marginal likelihood of each of the models using the methods of Kleibergen and van Dijk (1998).

[9] That model was $(0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0)$, where 0 means that the corresponding variable in Z^* entered in x_i rather than z_i. We take enough replications to ensure roughly 20000 iterations for each temperature.

[10] A run of 20000 iterations takes about one minute.

I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	Ex.	App.
1	1	1	1	1	1	1	1	1	1	1	1	1	0.42	0.41
1	1	1	1	1	1	1	1	1	1	0	1	1	0.30	0.31
1	1	1	1	0	1	1	1	1	1	1	1	1	0.07	0.07
1	1	1	1	1	1	1	1	0	1	1	1	1	0.06	0.05
1	1	1	1	0	1	1	1	0	1	1	1	1	0.04	0.04
1	1	1	1	1	1	1	1	0	1	0	1	1	0.02	0.02
1	1	1	1	1	1	1	1	1	0	1	1	1	0.01	0.01
1	1	1	1	1	1	1	1	1	0	0	1	1	0.01	0.01
1	0	1	1	1	1	1	1	1	1	1	1	1	0.01	0.01

Table 1: Posterior probability of best 9 models when T=0. The columns I1-I13 correspond to each of the 13 instruments. The value 1 indicates the potential instrument is included in z_i, and 0 indicates that it is included in x_i. The column labeled Ex. refers to the case in which each and every model of the whole model space is estimated separately. App. refers to the probabilities produced by the RJMCMC algorithm.

enters the model as an instrument. Again we can see that the posterior probabilities produced by our RJMCMC algorithm are very close to the ones produced by estimating each and every model. Note that again there are some differences between $T = 0$ and $T = 1$ (for instruments 5, 9 and 11) and $T = 2$ differs in that it has higher probability for these instruments. Our algorithm captures these differences.

I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	Ex.	App.
1	1	1	1	1	1	1	1	1	1	1	1	1	0.48	0.48
1	1	1	1	1	1	1	1	0	1	1	1	1	0.18	0.20
1	1	1	1	1	1	1	1	1	1	0	1	1	0.17	0.17
1	1	1	1	1	1	1	1	0	1	0	1	1	0.03	0.03
1	1	1	1	0	1	1	1	1	1	1	1	1	0.02	0.01
1	1	1	1	1	1	1	1	1	0	1	1	1	0.02	0.01
1	0	1	1	1	1	1	1	1	1	1	1	1	0.01	0.01
1	1	1	1	1	1	1	0	1	1	1	1	1	0.01	0.01
1	1	1	1	1	0	1	1	1	1	1	1	1	0.01	0.01

Table 2: Posterior probability of best 9 models when T=1.See Table 1 for definition of labels in columns.

I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	Ex.	App.
1	1	1	1	1	1	1	1	1	1	1	1	1	0.67	0.69
1	1	1	1	0	1	1	1	1	1	1	1	1	0.04	0.05
1	1	1	1	1	1	1	1	1	1	1	0	1	0.03	0.00
0	1	1	1	1	1	1	1	1	1	1	1	1	0.03	0.03
1	0	1	1	1	1	1	1	1	1	1	1	1	0.02	0.03
1	1	1	1	1	1	0	1	1	1	1	1	1	0.02	0.03
1	1	1	1	1	0	1	1	1	1	1	1	1	0.02	0.03
1	1	1	0	1	1	1	1	1	1	1	1	1	0.02	0.03
1	1	1	1	1	1	1	0	1	1	1	1	1	0.02	0.03
1	1	1	1	1	1	1	1	1	1	0	1	1	0.02	0.02
1	1	0	1	1	1	1	1	1	1	1	1	1	0.02	0.02

Table 3: Posterior probability of best models when T=2 (KvD prior). See Table 1 for definition of labels in columns. App. refers to estimating separately each of the T=1 models visited by the RJMCMC algorithm. Other labels defined as in Table 1.

	T=0		T=1		T=2	
Instrument	Ex.	App.	Ex.	App.	Ex.	App.
1	1.00	0.99	0.99	0.99	0.96	0.97
2	0.99	0.99	0.98	0.99	0.97	0.97
3	0.99	0.99	0.99	0.99	0.97	0.98
4	1.00	1.00	0.99	0.99	0.97	0.97
5	0.87	0.87	0.97	0.98	0.95	0.95
6	0.99	0.99	0.98	0.98	0.97	0.97
7	0.99	1.00	0.99	0.99	0.97	0.97
8	0.99	0.98	0.99	0.98	0.97	0.97
9	0.87	0.89	0.76	0.75	0.98	0.98
10	0.98	0.97	0.97	0.97	0.98	0.99
11	0.64	0.63	0.78	0.77	0.97	0.98
12	1.00	1.00	1.00	1.00	0.95	1.00
13	1.00	1.00	1.00	1.00	0.99	1.00

Table 4: Posterior Probability for each potential instrument of being included in z_i. T=2 refers to super-hot models that use the prior of Kleibergen and van Dijk (1998). The column labeled Ex. refers to the case in which the posterior probability is calculated by estimating each and every model of the whole model space. App. refers to the probabilities calculated by using the RJMCMC algorithm (T=0 and T=1). In the case of T=2, App. refers to estimating separately each of the models visited by the RJMCMC algorithm.

5.2 Empirical Results Using the Full Model Space

We now turn to the full model space and use the prior of Strachan and Inder (2004) for $T = 2$. We begin by presenting results which are not based on Bayesian model averaging. Table 5 gives frequentist and Bayesian estimates of the returns to schooling in the all-encompassing model (which includes all elements of X^* as exogenous regressors, all variables in Z^* as instruments, and treats all variables in y_2 as endogenous). The point estimates of the returns to schooling are similar using all approaches. The Bayesian posterior medians (for $T = 0, 1, 2$) lie in between the 2SLS (11%) and the LIML estimates (15%). The 95% Bayesian credible intervals for $T = 0, 1$ are wider than 2SLS confidence intervals but narrower than their LIML counterparts. However, the Bayesian credible interval with the prior $T = 2$ is much wider. It includes even negative values, indicating that identification might be poor[11]. The reason for the wider credible intervals with $T = 2$ is the non-existence of prior moments for γ.

In a BMA exercise it is typical to report not only averages over the whole model space, but also averages over restricted subspaces of the model space. For this purpose let us define four binary indicators: I_e, I_r, I_d and I_c which equal one for a particular subset of the models (the value zero indicates all of the relevant models are included). The indicator I_e takes value one when all variables in y_2 are endogenous (i.e. no exogeneity restrictions are imposed). I_r takes value 1 when the coefficients of (ED76, AGE76) are both different from 0.[12] I_d takes value one when $\widetilde{d}_u$ (the under-identification discrepancy measure) is above its own posterior median. Thus, I_d takes value one when identification is stronger. I_c takes value one when NEARC4 (conventionally considered to be a very important instrument) is included in the model as an instrument.

We begin with a discussion of returns to schooling estimates using BMA. These are given in Table 6. Results for BMA over the full model space are given in the rows labelled $T = 0, 1, 2$ in the column labelled $I_e = 0$. For all of our three temperatures, our point estimate of returns to schooling is 0.015 and 95% credible intervals are fairly narrow. This point estimate is

[11] The Anderson canonical correlation LR test rejects the null of under-identification (p-value 0.0066) while the Sargan test fails to reject the validity of over-identifying restrictions (p-value 0.2159). However, the Stock-Yogo test fails to reject that 2SLS estimates might be subject to 30% or more bias due to weak identification.

[12] The Data Appendix explains why this is an interesting restriction to consider.

substantively lower than those in Table 5. For instance, with LIML we found a point estimate of 0.146, and 0.015 is outside the LIML 95% confidence interval.

The other estimates in Table 6, using subsets of the model space, shed insight on why BMA is giving a lower estimate of returns to schooling than any of the other IV based approaches. Consider first what happens if we do BMA only over models in which all variables in y_2 are endogenous (i.e. $I_e = 1$). It can be seen that results are much more consistent with the non-BMA results of Table 5. That is, the 95% credible interval for $T = 0, 1$ exclude negative values and are centered at 10%. Just as we found in Table 5, the $T = 2$ credible interval is now very wide (and even contains negative values). If we consider additional restrictions on the model space, the basic story (i.e. that only considering models where all variables in y_2 are endogenous is necessary to obtain results similar to those found using standard IV methods) is not altered. That is, conditioning on $(I_r = 1, I_d = 1)$ or $(I_r = 1, I_c = 1)$ does not change results in a substantive fashion.

Table 7 shows the prior and posterior percentiles of $\widetilde{d}_u$ and this gives evidence that identification is much weaker when $I_e = 1$.[13] This is a point we will return to shortly.

Tables 8, 9 and 10 show the probability that each variable enters each category. Tables 8 and 10 indicate that BMA has a strong preference for parsimony. Our full model space allows the elements of Z^* to enter as instruments, as exogenous regressors or be excluded from the model. Tables 8 and 10 indicate that some are included as instruments, but most are excluded altogether from the model. Similarly, BMA allows the variables in X^* (which were always included in the models used to produce the results in Table 4) to be either exogenous regressors or be excluded from the model. Table 10 indicates most are excluded from the model. Table 9 provides us with strong evidence that two of the three elements of our "endogenous" y_2 are actually exogenous. And Figure 1, which shows BMA posterior densities (conditional on $I_e = 1$) of the correlations between v_2 and u_1, supports this view. Lastly, the 2SLS estimate of the returns to schooling in the best model selected[14] by

[13] The discrepancy measure $\widetilde{d}_o$ is not shown in some cases because there is probability one of just-identification for each temperature, with the unused instruments not entering in the model at all. Two of these three instruments are AGE762 and IQ with probability 1. The third one could be either EDFDUM2 or EDFDUM1.

[14] By best model we mean the model that results from rounding the posterior probability of each restriction to the nearest integer (0 or 1).

OLS	0.051	0.061	0.072
2SLS	0.033	0.108	0.183
LIML	0.035	0.146	0.256
T=0	0.047	0.123	0.206
T=1	0.030	0.113	0.206
T=2	-0.075	0.140	0.517

Table 5: Frequentist and Bayesian estimates and 95% confidence intervals of returns to schooling in the all encompassing model.

the RJMCMC is 0.012, with 95% confidence interval being (0.004, 0.021).

Putting all these findings together, we can now see why BMA is estimating returns to schooling as being lower than the traditional IV approaches of Table 5. Most importantly, the assumption that the elements of y_2 truly are endogenous is crucial to obtaining the traditional IV results. However, BMA is allocating relatively little weight to such models. Averaging over the full model space (i.e. including also models with exogeneity restrictions imposed) helps identification and makes credible intervals of the returns to schooling narrower and centered on 1.5% for each of the 3 temperatures (Table 6). The posterior of $\widetilde{d}_u$ confirms that identification is substantially stronger if we use the full model space, and $\widetilde{d}_o$ shows that the over-identifying restrictions hold (Table 7). The probability that only three elements of Z^* enter as instruments is 100% for $T = 0, 1, 2$. The most likely instruments are AGE762 and IQ, followed by EFDUM1 and EFDUM2. A further difference between BMA and non-BMA results arises since the former is much more parsimonious than the latter (and this holds for all of our priors).

In sum, this empirical example shows that our RJMCMC algorithm can be used to carry out BMA even in the very large model spaces that the researcher will often encounter in practice. It also shows that BMA can matter empirically. That is, BMA is leading to estimates of a feature of interest (returns to schooling) which differ in important ways from conventional estimates. Furthermore, it provides insight into why such divergences occur and what aspects of model specification have the most important impact on estimates of the returns to schooling.

	I_e=0			I_e=1		
T=0	0.004	0.015	0.088	0.015	0.099	0.120
T=0, I_r=1	0.043	0.085	0.127	0.087	0.105	0.123
T=0, I_r=1, I_d=1	0.003	0.069	0.089	0.087	0.105	0.123
T=0, I_r=1, I_c=1	0.069	0.085	0.126	0.088	0.105	0.124
T=1	0.005	0.017	0.088	0.015	0.099	0.120
T=1, I_r=1	0.040	0.084	0.126	0.087	0.105	0.123
T=1, I_r=1, I_d=1	0.004	0.068	0.088	0.087	0.105	0.123
T=1, I_r=1, I_c=1	0.067	0.084	0.125	0.088	0.105	0.124
T=2	0.005	0.014	0.024	-0.022	0.019	0.118
T=2, I_r=1	0.006	0.036	0.125	-0.291	0.104	0.523
T=2, I_r=1, I_d=1	-0.002	0.025	0.048	-0.166	0.103	0.260
T=2, I_r=1, I_c=1	0.003	0.106	0.129	-0.998	0.108	1.338

Table 6: BMA posterior percentiles (2.5%, 50%, 97.5%) of returns to schooling. The columns under I_e=1 correspond to the case in which exogeneity restrictions are not considered, while those under I_e=0 refer to the case in which the model space includes also models with exogeneity restrictions.

		I_e=0			I_e=1		
$\tilde{d}_u$	T=0 prior	1.11	299901	2460014	0.000	0.001	0.318
	T=0 post	0.001	3.59	8.12	0.000	0.001	0.002
	T=1 prior	1.35	270274	2367542	0.000	0.001	0.320
	T=1 post	0.001	2.58	7.13	0.000	0.001	0.002
	T=2 prior	0.047	435	3117025	0.000	0.002	0.542
	T=2 post	0.001	0.401	7.35	0.000	0.001	0.002
$\tilde{d}_o$	T=0 prior	0.015	201	6223			
	T=0 post	0.000	0.000	0.005			
	T=1 prior	0.017	132	5897			
	T=1 post	0.000	0.000	0.005			
	T=2 prior	0.001	8.70	5482			
	T=2 post	0.000	0.000	0.004			

Table 7: BMA posterior percentiles (2.5%, 50%, 97.5%) of ($\tilde{d}_o$, $\tilde{d}_u$).

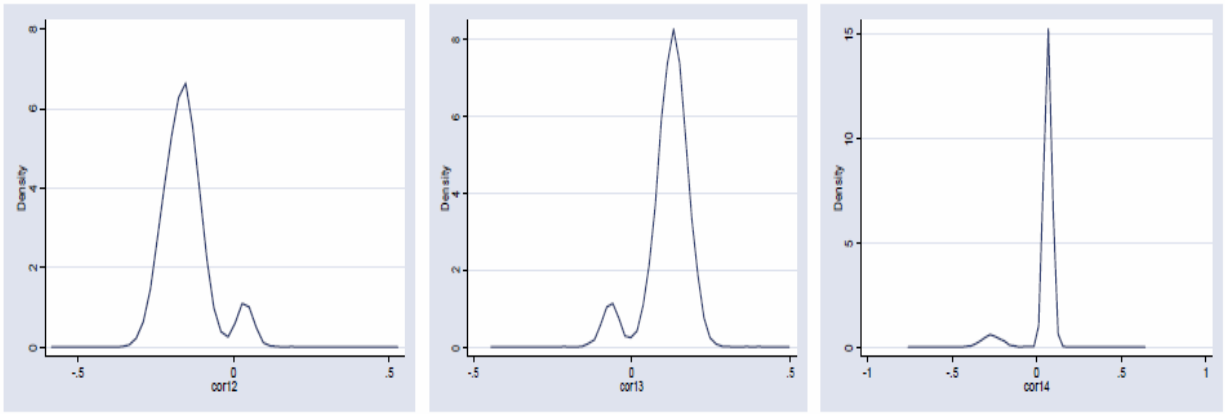

Figure 1: Figure 1: Posterior density for the correlation between u and v conditional on $I_e = 1$ and $T = 1$. From left to right the correlations correspond to u and the error terms of ED76, EXPER2 and KWW.

	I_e=0			I_e=1		
	T=0	T=1	T=2	T=0	T=1	T=2
EDFDUM1	0.63	0.61	0.76	0.08	0.09	0.39
EDFDUM2	0.37	0.39	0.24	0.92	0.91	0.61
EDFDUM3	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM4	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM5	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM6	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM7	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM8	0.00	0.00	0.00	0.00	0.00	0.00
NEARC4	0.00	0.00	0.00	0.00	0.00	0.00
NEARC2	0.00	0.00	0.00	0.00	0.00	0.00
NEARC4A	0.00	0.00	0.00	0.00	0.00	0.00
AGE762	1.00	1.00	1.00	1.00	1.00	1.00
IQ	1.00	1.00	1.00	1.00	1.00	1.00

Table 8: Probability of variables in Z^* entering in the model as an instrument (in z).

	I_e = 0		
	T=0	T=1	T=2
ED76	0.05	0.04	0.05
EXPER2	0.07	0.06	0.47
KWW	0.98	0.98	1.00

Table 9: Probabilities of ED76, EXPER2 and KWW being endogenous.

	I_e=0			I_e=1		
	T=0	T=1	T=2	T=0	T=1	T=2
ED76	0.03	0.04	0.00	0.45	0.46	0.10
EXPER2	0.28	0.37	0.01	0.49	0.50	0.16
KWW	0.70	0.62	1.00	0.09	0.09	0.95
AGE76	1.00	1.00	1.00	1.00	1.00	1.00
BLACK	1.00	1.00	1.00	1.00	1.00	1.00
SMSA76R	1.00	1.00	1.00	1.00	1.00	1.00
REG76R	0.00	0.00	0.00	0.00	0.00	0.00
FDUM1	0.00	0.00	0.00	0.00	0.00	0.00
FDUM2	0.00	0.00	0.00	0.00	0.00	0.00
FDUM3	0.00	0.00	0.00	0.00	0.00	0.00
FDUM4	0.00	0.00	0.00	0.00	0.00	0.00
FDUM5	0.00	0.00	0.00	0.00	0.00	0.00
FDUM6	0.00	0.00	0.00	0.00	0.00	0.00
FDUM7	0.00	0.00	0.00	0.00	0.00	0.00
FDUM8	0.00	0.00	0.00	0.00	0.00	0.00
REG661	0.00	0.00	0.00	0.00	0.00	0.00
REG662	0.00	0.00	0.00	0.00	0.00	0.00
REG663	0.00	0.00	0.00	0.00	0.00	0.00
REG664	0.00	0.00	0.00	0.00	0.00	0.00
REG665	0.00	0.00	0.00	0.00	0.00	0.00
REG666	0.00	0.00	0.00	0.00	0.00	0.00
REG667	0.00	0.00	0.00	0.00	0.00	0.00
REG668	0.00	0.00	0.00	0.00	0.00	0.00
SMSA66R	0.00	0.00	0.00	0.00	0.00	0.00
MOMDAD14	0.00	0.00	0.00	0.00	0.00	0.00
SINMOM14	0.00	0.00	0.00	0.00	0.00	0.00
DADED	1.00	1.00	1.00	1.00	1.00	1.00
MOMED	0.00	0.00	0.00	0.00	0.00	0.00
NODADED	0.00	0.00	0.00	0.00	0.00	0.00
NOMOMED	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM1	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM2	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM3	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM4	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM5	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM6	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM7	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM8	0.00	0.00	0.00	0.00	0.00	0.00
NEARC4	0.00	0.00	0.00	0.00	0.00	0.00
NEARC2	0.00	0.00	0.00	0.00	0.00	0.00
NEARC4A	0.00	0.00	0.00	0.00	0.00	0.00
AGE762	0.00	0.00	0.00	0.00	0.00	0.00
IQ	0.00	0.00	0.00	0.00	0.00	0.00

Table 10: Probability of being included as a regressor in the first structural equation (in x or y_2).

6 Conclusions

BMA has enjoyed an increasing popularity amongst econometricians working with the regression model with a large number of exogenous regressors. The purpose of the present paper is to develop methods for BMA when endogeneity may be present. In such a case, any variable could be an endogenous variable, an exogenous variable or an instrument (and sometimes the researcher is unsure which category a variable belongs to). Doing BMA with such a setup is complicated by the huge model space that results and (in contrast to the case where all regressors are exogenous) the lack of availability of analytical results for each model. To surmount these problems, this paper develops a RJMCMC algorithm which draws jointly from the model and parameter spaces. To surmount problems of slow convergence, we draw on ideas from the simulated tempering literature and introduce cold, hot and super-hot models into our algorithm. A further advantage of our algorithm is that draws of different temperatures can be used to carry out Bayesian inference under different priors. If we use the draws from the cold models we are doing BMA under an approximation to the posterior, if we use hot draws we are doing BMA using the prior of Drèze (1976) and if we use super-hot draws we are doing BMA using a prior such as that of Strachan and Inder (2004).

We illustrate our algorithm using the classic returns to schooling application of Card (1995). We find our RJMCMC algorithm to work efficiently and empirical results show some interesting differences between model averaging and conventional econometric methodologies.

References

Anderson, T. 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.

Anderson, T. and Rubin, H., 1950, The asymptotic properties of estimators of the parameters of a single equation in a complete system of stochastic equations, *Annals of Mathematical Statistics,* 21, 570-582.

Bauwens, L., Lubrano, M. and Richard, J.-F., 1999, *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Card, D., 1995, Using geographic variation in college proximity to estimate the return to schooling, in *Aspects of Labour Market Behaviour: Essays in Honour of John Vandekamp* edited by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press.

Cohen-Cole, E., Durlauf, S., Fagan, J. and Nagin, D., 2009, Model uncertainty and the deterrent effect of capital punishment, *American Law and Economics Review*, forthcoming.

Drèze, J.H., 1976, Bayesian limited information analysis of the simultaneous equations model, *Econometrica*, 44, 1045–1075.

Drèze, J.H. and Richard, J.F. 1983, Bayesian Analysis of Simultaneous Equations Systems. In *Handbook of Econometrics*, volume 1, edited by Z. Griliches and M.D. Intrilligator. Amsterdam: Elsevier Science.

Eicher, T.S., Lenkoski, A. and Raftery, A.E. (2009), Bayesian model averaging and endogeneity under model uncertainty: An application to development determinants, *Working Paper no. 94*, Center for Statistics and the Social Sciences, University of Washington.

Fernandez, C., Ley, E. and Steel, M., 2001, Benchmark priors for Bayesian model averaging, *Journal of Econometrics*, 100, 381-427.

Geyer, C. and Thompson, E., 1995, Annealing Markov Chain Monte Carlo with applications to ancestral inference, *Journal of the American Statistical Association,* 90, 909-920.

Green, P., 1995, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82, 711-732.

Greene, W., 2003, *Econometric Analysis* (Fifth edition), New Jersey: Prentice-Hall.

Hausman, J., 1983, Specification and estimation of simultaneous equations models, in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, volume 1. Amsterdam: North Holland.

Holmes, C., Denison, D. and Mallick, B., 2002, Bayesian model order determination and basis selection for seemingly unrelated regression, *Journal of Computational and Graphical Statistics*, 11, 533–551.

Johansen, S., 1988, Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics and Control*, 12, 231-254.

Johansen, S., 1995, *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Kleibergen, F. and Paap, R., 2002, Priors, posteriors and bayes factors for a Bayesian analysis of cointegration, *Journal of Econometrics*, 111, 223-249.

Kleibergen, F. and van Dijk, H., 1998, Bayesian simultaneous equations analysis using reduced rank structures, *Econometric Theory*, 14, 699-744.

Koop, G., 2003, *Bayesian Econometrics.* Chichester: Wiley.

Koop, G., Leon-Gonzalez, R. and R. Strachan, 2010, Efficient posterior simulation for cointegrated models with priors on the cointegration space, *Econometric Reviews*, 29, 224-242.

Liu, J.S., 2001, *Monte Carlo Strategies in Scientific Computing.* Berlin: Springer.

Madigan, D. and York, J., 1995, Bayesian graphical models for discrete data, *International Statistical Review*, 63, 215-232.

Madan, D.B. and Seneta, E., 1990, The Variance-Gamma (V.G) Model for Share Market Returns, *Journal of Business* 63, 511-524.

Marinari, E. and Parisi, G., 1992, Simulated tempering: a new Monte Carlo scheme., *Europhysics Letters*, 19, 451-458.

Muirhead, R.J., 1982, *Aspects of Multivariate Statistical Theory.* New York: Wiley.

Sargan, J., 1958, The estimation of economic relationships using instrumental variables, *Econometrica*, 26, 393-415.

Strachan, R. and Inder, B., 2004, Bayesian analysis of the error correction model, *Journal of Econometrics*, 123, 307-325.

Waagepetersen, R. and Sorensen, D., 2001, A tutorial on reversible jump MCMC with a view toward applications in QLT mapping, *International Statistical Review*, 69, 49-61.

Zellner, A., 1971, *An Introduction to Bayesian Inference in Econometrics.* New York: John Wiley and Sons.

Zellner, A., Bauwens, L. and van Dijk, H.K., 1988, Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods, *Journal of Econometrics*, 38, 39-72.

Data Appendix

The data used in this paper was used in Card (1995) and provided on Card's website: http://emlab.berkeley.edu/users/card/data_sets.html. These sources provide complete information about this data set. We use $N = 2040$ observations on individuals from 1976 from the National Longitudinal Survey (this is the original cohort). In our modelling approach, each variable must either be the main dependent variable of interest (y_1), another endogenous variable (y_2), a potential regressor (X^*) or a variable which could either be an instrument or a regressor (Z^*). We follow Card (1995) in our classification of variables and refer the reader to his paper for a justification. The following is a summary of the 45 variables we use along with the category each belongs in. All variables refer to 1976 unless otherwise noted.

Table A.1: Variables used in Application		
Name	Brief Description	Type
LWAGE76	log wages	y_1
ED76	education	y_2
EXPER2[15]	experience squared/100	y_2
KWW	score on Knowledge of World of Work test	y_2
AGE76	Age	X^*
BLACK	Dummy for black	X^*
SMSA76R	Dummy for urban	X^*
REG76R	Dummy for south	X^*
FDUM1	Mom and Dad both >12 years education	X^*
FDUM2	Mom and Dad ≥12 and not both exactly 12	X^*
FDUM3	Mom and Dad both =12 years education	X^*
FDUM4	Mom ≥12 years education and Dad missing	X^*
FDUM5	Dad≥12 and Mom<12 years education	X^*
FDUM6	Mom ≥12 years education and Dad non-missing	X^*
FDUM7	Mom and Dad both ≥9 years education	X^*
FDUM8	Mom and Dad both non-missing	X^*
REG661	Dummy for region 1 in 1966	X^*
REG662	Dummy for region 2 in 1966	X^*
REG663	Dummy for region 3 in 1966	X^*
REG664	Dummy for region 4 in 1966	X^*
REG665	Dummy for region 5 in 1966	X^*
REG666	Dummy for region 6 in 1966	X^*
REG667	Dummy for region 7 in 1966	X^*
REG668	Dummy for region 8 in 1966	X^*

[15] Card defines experience as age - education - 6 and includes it, together with EXPER2, as an endogenous explanatory variable while age is included as an instrument. To avoid having a singular covariance matrix, we instead include age as a regressor (i.e. in X^*) and exclude experience from the analysis (but still include EXPER2 in y_2). Note that our specification is just a reparameterization of that of Card (1995), and in our case the return to schooling is given by the sum of the coefficients of ED76 and AGE76.

Table A.1 (continued): Variables used in Application		
Name	Brief Description	Type
SMSA66R	Dummy for urban in 1966	X^*
MOMDAD14	Dummy for living with mom and dad at 14	X^*
SINMOM14	Dummy for living with single mom at 14	X^*
DADED	Dad's years of schooling	X^*
MOMED	Mom's years of schooling	X^*
NODADED	Dummy for DADED imputed	X^*
NOMOMED	Dummy for MOMED imputed	X^*
EDFDUM1	FDUM1*NEARC4	Z^*
EDFDUM2	FDUM2*NEARC4	Z^*
EDFDUM3	FDUM3*NEARC4	Z^*
EDFDUM4	FDUM4*NEARC4	Z^*
EDFDUM5	FDUM5*NEARC4	Z^*
EDFDUM6	FDUM6*NEARC4	Z^*
EDFDUM7	FDUM7*NEARC4	Z^*
EDFDUM8	FDUM8*NEARC4	Z^*
NEARC4	Dummy grew up near any 4 year college	Z^*
NEARC2	Dummy grew up near 2 year college	Z^*
NEARC4A	Dummy grew up near 4 year public college	Z^*
AGE762	Age squared	Z^*
IQ	Normed IQ score	Z^*

Technical Appendix

Algorithm

To illustrate the general principle underlying the algorithm we use, suppose that the vector of unknown parameters in model M can be decomposed as $\theta_M = (\theta_{1M}, \theta_{2M})$. Let $q(M^{(*)}|M^{(r)})$ be a proposal density for models. Because we are going to define a move conditional on θ_{1M}, we require that $q(M^{(*)}|M^{(r)})$ gives zero probability to models $M^{(*)}$ in which the dimension of θ_{1M} changes. Let $q(\theta_{2M}|\theta_{1M}, M)$ be a proposal density for θ_{2M}. The general expression for the acceptance probability for a move from $(\theta^{(r)}_{2M^{(r)}}, M^{(r)})$ to $(\theta^{(*)}_{2M^{(*)}}, M^{(*)})$ conditional on θ_{1M} can be found for example at Waagepetersen and Sorensen (2001) and it is equal to:

$$a = \min\left\{1, \frac{q(M^{(r)}|M^{(*)})}{q(M^{(*)}|M^{(r)})} \frac{p(Y, \theta_{1M}, \theta^{(*)}_{2M^{(*)}}|M^{(*)})}{p(Y, \theta_{1M}, \theta^{(r)}_{2M^{(r)}}|M^{(r)})} \frac{q(\theta^{(r)}_{2M^{(r)}}|\theta_{1M}, M^{(r)})}{q(\theta^{(*)}_{2M^{(*)}}|\theta_{1M}, M^{(*)})} \frac{p(M^{(*)})}{p(M^{(r)})}\right\}$$

where $p(M^*)$ is the prior probability of model M^*. Following the strategy of Holmes and Held (2006), we always choose $q(\theta^{(*)}_{2M^{(*)}}|\theta_{1M}, M^{(*)})$ to be the optimal choice $p(\theta^{(*)}_{2M^{(*)}}|Y, \theta_{1M}, M^{(*)})$, that is, the conditional posterior of θ_{2M} given θ_{1M} and $M = M^{(*)}$. As a consequence of choosing such proposal density, the expression for a simplifies to:

$$a = \min\left\{1, \frac{q(M^{(r)}|M^{(*)})}{q(M^{(*)}|M^{(r)})} \frac{p(Y, \theta_{1M}|M^{(*)})}{p(Y, \theta_{1M}|M^{(r)})} \frac{p(M^{(*)})}{p(M^{(n)})}\right\} \tag{7}$$

where

$$p(Y, \theta_{1M}|M) = \int p(Y, \theta_{1M}, \theta_{2M}|M) d\theta_{2M} = \int p(\theta_{1M}, \theta_{2M}|M) p(Y|\theta_{1M}, \theta_{2M}, M) d\theta_{2M}$$

We use two indexes to describe the model space: (M, T), where T takes values 0 (for cold models, which are based on an approximation to the posterior), 1 (for hot models, which use Drèze's prior) and 2 (for super-hot models, which use another prior $p^*(\Pi, \Sigma|M)$, where $\Pi = (\beta', \gamma', vec(\Pi_{2x}), vec(\Pi_{2z})')$. Let the prior probability of each (M, T) be denoted as $p(M, T) = p(T)p(M|T)$. The function $p(T)$ can be chosen as a tuning parameter to ensure that the algorithm spends enough time at each temperature. Let $(\Pi^{(r)}, \Sigma^{(r)}, M^{(r)}, T^{(r)})$

be the value of (Π, Σ, M, T) in the r^{th} draw from the algorithm. Our proposal density for (M, T), which we denote as $q(M^{(*)}, T^{(*)}|M^{(r)}, T^{(r)})$, is such that with probability $\rho_{T^{(r)}}$ a candidate value for temperature $(T^{(*)})$ is drawn from some distribution $(q(T^{(*)}|T^{(r)}))$ while the model restrictions remain constant (i.e. $M^{(*)} = M^{(r)}$) and with probability $(1 - \rho_{T^{(r)}})$ a candidate model $(M^{(*)})$ is drawn from some distribution $(q(M^{(*)}|M^{(r)}, T^{(r)}))$ while the value of temperature remains constant $(T^{(*)} = T^{(r)})$. The values defining $\rho_{T^{(r)}}$ are denoted as τ_1^* and τ_2^*, with $\tau_1^* \leq \tau_2^*$. These are constants that, together with $p(T)$, can be calibrated in the burn-in period to ensure that the algorithm visits each temperature enough times[16].

The $(r+1)^{th}$ value of (Π, Σ, M, T) (denoted as $(\Pi^{(r+1)}, \Sigma^{(r+1)}, M^{(r+1)}, T^{(r+1)})$ is obtained as follows:

If $T^{(r)} = 0$:

- Draw u from a uniform in $(0, 1)$.
- If $u < \tau_1^*$:(propose a change from a cold model to the analogous hot one conditioning only on Π_{2z}).
  - Fix $M^{(r+1)} = M^{(r)}$. Fix $T^{(r+1)} = 1$ with probability a and fix $T^{(r+1)} = 0$ with probability $(1 - a)$, where a is defined as:

$$a = \min \left\{ \frac{p(M^{(r+1)}, T^{(r+1)} = 1) p(Y, \Pi_{2z}^{(r)} | M^{(r+1)}, T^{(r+1)} = 1)}{p(M^{(r+1)}, T^{(r+1)} = 0) p(Y, \Pi_{2z}^{(r)} | M^{(r+1)}, T^{(r+1)} = 0)}, 1 \right\}$$

  - If $T^{(r+1)} = 1$ draw $\Sigma^{(r+1)}$ conditional on $(\Pi_{2z}^{(r)}, M^{(r+1)}, T^{(r+1)})$ and then draw $(\Pi_{2z}^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \Pi_{2x}^{(r+1)})$ conditional on $(\Sigma^{(r+1)}, M^{(r+1)}, T^{(r+1)})$.
  - If $T^{(r+1)} = 0$ draw $(\Pi^{(r+1)}, \Sigma^{(r+1)})$ conditional on $(M^{(r+1)}, T^{(r+1)})$.
- If $u \geq \tau_1^*$: (propose a change from a cold model to another cold model, changing any of the model restrictions)
  - Fix $T^{(r+1)} = T^{(r)} = 0$. Draw a candidate value $M^{(*)}$ from a proposal distribution $q(M|M^{(r)}, T^{(r+1)} = 0)$. This proposal distribution changes any of the model restrictions with some probability.

[16] Liu (2001, p. 210) recommends that simulated tempering algorithms are tuned so that all temperatures are visited with the same frequency.

Fix $M^{(r+1)} = M^{(*)}$ with probability a and fix $M^{(r+1)} = M^{(r)}$ with probability $(1-a)$, where a is defined as:

$$a = \min\left\{\frac{p(M^{(*)}, T^{(r+1)})p(Y|M^{(*)}, T^{(r+1)})q(M^{(r)}|M^{(*)}, T^{(r+1)})}{p(M^{(r)}, T^{(r+1)})p(Y|M^{(r)}, T^{(r+1)})q(M^{(*)}|M^{(r)}, T^{(r+1)})}, 1\right\}$$

– Draw $(\Pi^{(r+1)}, \Sigma^{(r+1)})$ conditional on $(M^{(r+1)}, T^{(r+1)})$.

If $T^{(r)} = 1$:

- Draw u from a uniform in $(0, 1)$.
- If $u < \tau_1^*$: (propose a change from a hot model to the analogous cold one conditioning only on Π_{2z}).
  - Fix $M^{(r+1)} = M^{(r)}$. Fix $T^{(r+1)} = 1$ with probability a and fix $T^{(r+1)} = 0$ with probability $(1-a)$, where a is defined as:

$$a = \min\left\{\frac{p(M^{(r+1)}, T^{(r+1)} = 0)p(Y, \Pi_{2z}^{(r)}|M^{(r+1)}, T^{(r+1)} = 0)}{p(M^{(r+1)}, T^{(r+1)} = 1)p(Y, \Pi_{2z}^{(r)}|M^{(r+1)}, T^{(r+1)} = 1)}, 1\right\}$$

  - If $T^{(r+1)} = 1$ draw $\Sigma^{(r+1)}$ conditional on $(\Pi_{2z}^{(r)}, M^{(r+1)}, T^{(r+1)})$ and then draw $(\Pi_{2z}^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \Pi_{2x}^{(r+1)})$ conditional on $(\Sigma^{(r+1)}, M^{(r+1)}, T^{(r+1)})$.
  - If $T^{(r+1)} = 0$ draw $(\Pi^{(r+1)}, \Sigma^{(r+1)})$ conditional on $(M^{(r+1)}, T^{(r+1)})$.
- If $\tau_1^* \leq u \leq \tau_2^*$: (propose a change from a hot model to another hot model conditioning on Σ).
  - Fix $T^{(r+1)} = T^{(r)} = 1$. Draw a candidate value $M^{(*)}$ from a proposal distribution $q(M|M^{(r)}, T^{(r+1)})$. This distribution proposes models that could change any restriction except for those related to Σ. Fix $M^{(r+1)} = M^{(*)}$ with probability a and fix $M^{(r+1)} = M^{(r)}$ with probability $(1-a)$, where a is defined as:

$$a = \min\left\{\frac{p(M^{(*)}, T^{(r+1)})p(Y, \Sigma^{(r)}|M^{(*)}, T^{(r+1)})q(M^{(r)}|M^{(*)}, T^{(r+1)})}{p(M^{(r)}, T^{(r+1)})p(Y, \Sigma^{(r)}|M^{(r)}, T^{(r+1)})q(M^{(*)}|M^{(r)}, T^{(r+1)})}, 1\right\}$$

- Draw $\Pi_{2z}^{(r+1)}$ conditional on $(\Sigma^{(r)}, M^{(r+1)}, T^{(r+1)})$ and then draw $(\Sigma^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \Pi_{2x}^{(r+1)})$ conditional on $(\Pi_{2z}^{(r+1)}, M^{(r+1)}, T^{(r+1)})$..

- If $\tau_2^* \leq u$: (propose a change from a hot model to the analogous super-hot model conditioning on all parameters):

  - Fix $M^{(r+1)} = M^{(r)}$. Fix $T^{(r+1)} = 2$ with probability a and fix $T^{(r+1)} = 1$ with probability $(1-a)$, where a is defined as the minimum of 1 and:

$$(1-\tau_2^*)\frac{p(M^{(r+1)}, T^{(r+1)}=2)p(\Pi^{(r)}, \Sigma^{(r)}|M^{(r+1)}, T^{(r+1)}=2)}{p(M^{(r+1)}, T^{(r+1)}=1)p(\Pi^{(r)}, \Sigma^{(r)}|M^{(r+1)}, T^{(r+1)}=1)}$$

  - If $T^{(r+1)} = 1$: Draw $\Sigma^{(r+1)}$ conditional on $(\Pi_{2z}^{(r)}, M^{(r+1)}, T^{(r+1)})$ and then draw $(\Pi_{2z}^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \Pi_{2x}^{(r+1)})$ conditional on $(\Sigma^{(r+1)}, M^{(r+1)}, T^{(r+1)})$.
  - If $T^{(r+1)} = 2$: Draw $(\Pi^{(r+1)}, \Sigma^{(r+1)})$ conditional on $(M^{(r+1)}, T^{(r+1)} = 2)$ using a kernel $P^*(\Pi^{(r+1)}, \Sigma^{(r+1)}|\Pi, \Sigma)$ that is invariant for the posterior $p(\Pi^{(r+1)}, \Sigma^{(r+1)}|Y, M^{(r+1)}, T^{(r+1)} = 2)$.

If $T^{(r)} = 2$:

- (Propose a change from a super-hot model to the analogous hot model conditioning on all parameters):

  - Fix $M^{(r+1)} = M^{(r)}$. Fix $T^{(r+1)} = 1$ with probability a and fix $T^{(r+1)} = 2$ with probability $(1-a)$, where a is defined as the minimum of 1 and:

$$\frac{1}{(1-\tau_2^*)}\frac{p(M^{(r+1)}, T^{(r+1)}=1)p(\Pi^{(r)}, \Sigma^{(r)}|M^{(r+1)}, T^{(r+1)}=1)}{p(M^{(r+1)}, T^{(r+1)}=2)p(\Pi^{(r)}, \Sigma^{(r)}|M^{(r+1)}, T^{(r+1)}=2)}$$

  - If $T^{(r+1)} = 1$: Draw $\Sigma^{(r+1)}$ conditional on $(\Pi_{2z}^{(r)}, M^{(r+1)}, T^{(r+1)})$ and then draw $(\Pi_{2z}^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \Pi_{2x}^{(r+1)})$ conditional on $(\Sigma^{(r+1)}, M^{(r+1)}, T^{(r+1)})$.
  - If $T^{(r+1)} = 2$: Draw $(\Pi^{(r+1)}, \Sigma^{(r+1)})$ conditional on $(M^{(r+1)}, T^{(r+1)} = 2)$ using a kernel $P^*(\Pi^{(r+1)}, \Sigma^{(r+1)}|\Pi, \Sigma)$ that is invariant for the posterior $p(\Pi^{(r+1)}, \Sigma^{(r+1)}|Y, M^{(r+1)}, T^{(r+1)} = 2)$.

Note that when we use the ratio of priors $p(\Pi, \Sigma|T = 2)/p(\Pi, \Sigma|T = 1)$, both priors must use the same parameterization (i.e. Π, Σ). Therefore for most priors we will have to use the Jacobian of the transformation in order to write $p(.|T = 2)$ using the same parameterization as $p(.|T = 1)$. We give the relevant Jacobian for the Strachan and Inder (2004) type prior below. However, as we use a parameter-augmented version of the prior of Strachan and Inder (2004), this implies that $p(.|T = 2)$ will not only depend on (Π, Σ), but also on an additional non-identified matrix that we will denote as α_2. To deal with this, we augment also the Drèze prior with the additional parameter α_2, and so define $p(\Pi, \Sigma, \alpha_2|T = 1) = p(\Pi, \Sigma|T = 1)\varpi(\alpha_2)$. The density $\varpi(\alpha_2)$ could in principle be any, but we choose it to be equal to the marginal prior of α_2 in the setup described below. In this way, the ratio of priors entering in the acceptance probability will be $p(\Pi, \Sigma, \alpha_2|M, T = 2)/p(\Pi, \Sigma, \alpha_2|M, T = 1)$.

The proposal density for models ($q(M^*|M, T)$) could be any provided that it satisfies the following requirement: any model in the model space could be proposed with some positive probability after a finite number of iterations. In order to describe the proposal density that we use let us define 5 binary vectors (η_1,η_2,η_3,η_4, η_E) that determine the restrictions in a model. The binary vector η_1 has as many elements as potential regressors there are in X^*. It takes value 1 when the corresponding regressor enters in x_i, and value 0 when it is excluded from the system. The vector η_4 corresponds to y_2. It takes value 1 when the corresponding element in γ is non-zero. η_E has also as many elements as y_2. It takes value 1 when the corresponding variable is endogenous (i.e. the corresponding element of σ_{12} is non-zero). Each of the vectors η_2 and η_3 has as many elements as potential instruments there are in Z^*. An element in η_2 is 0 when the corresponding variable in Z^* is excluded from the system and takes value 1 when it is included (either in x_i or in z_i). Finally, an element in η_3 is 1 when the corresponding variable in Z^* enters in z_i, and takes value 0 otherwise. Note that if an element in $\eta_3^{(r)}$ is one, then the corresponding element in $\eta_2^{(r)}$ must also be one. Thus the current model $M^{(r)}$ can be described by the 5 binary vectors $(\eta_1^{(r)}, \eta_2^{(r)}, \eta_3^{(r)}, \eta_4^{(r)}, \eta_E^{(r)})$.

For the current model $M^{(r)}$ let us consider 4 types of movements: $C = 1$) Change only (η_1, η_2, η_3), $C = 2$) Change only η_3, $C = 3$) Change only η_4, $C = 4$) Change only η_E. Under $T = 0$, $q(M^*|M, T = 0)$ is such that C can take values $(1, 2, 3, 4)$ each with equal probability $(1/4)$. However, when $T = 1$, C takes only values (1,2,3), each with probability 1/3. Conditional

on $C = 4$, one of the elements in $\eta_E^{(r)}$ is chosen randomly and its current value is changed (from 0 to 1 if the current value is 0 or otherwise from 1 to 0). Similarly, conditional on $C = 3$ one of the elements in $\eta_4^{(r)}$ is chosen randomly and changed.

When $C = 1$ we will propose movements that take instruments/regressors in and out of the model. Let $\widetilde{\eta}_2^{(r)}$ be those elements of $\eta_2^{(r)}$ that correspond to potential instruments that are currently not in x_i (that is, those potential instruments are either out of the model ($\eta_2^{(r)} = 0$) or in z_i ($\eta_2^{(r)} = 1, \eta_3^{(r)} = 1$)). With the intention of improving convergence speed, when $C = 1$ we do not only consider increasing or decreasing the number of instruments/regressors by just one, but also we allow for a move that changes the set of instruments while keeping the number of instruments the same. That is, with probability ç one of the elements in $(\eta_1^{(r)\prime}, \widetilde{\eta}_2^{(r)\prime})'$ is chosen randomly and its current value is changed. This is a move that changes the number of instruments/regressors. If an element of $\eta_1^{(r)}$ is chosen, only its own value will be changed. But if an element of $\widetilde{\eta}_2^{(r)}$ is chosen, the corresponding value in $\eta_2^{(r)}$ **and** in $\eta_3^{(r)}$ will be changed. For example, if a variable in Z^* was out of the model and is chosen, it will be proposed as an instrument (in z_i). But if it was already in z_i, the proposed movement will take it out from the model. Let $Z_{-x}^{*(r)}$ be the set of potential instruments that are currently not in x_i and let $\widetilde{\eta}_3^{(r)}$ be all elements of $\eta_3^{(r)}$ except for those that are currently in x_i. If $Z_{-x}^{*(r)}$ is not an empty set, with probability (1-ç) we change a random number of elements in $\widetilde{\eta}_3^{(r)}$ while leaving the value $(\widetilde{\eta}_3^{(r)\prime}\widetilde{\eta}_3^{(r)})$ constant (however if $Z_{-x}^{*(r)}$ is the empty set, with probability (1-ç) the candidate model will be equal to the current one). The instruments to replace the current ones are going to be chosen from $Z_{-x}^{*(r)}$. To do this, the number of elements of $\widetilde{\eta}_3^{(r)}$ to be changed, denote it as $\hbar$, is drawn from a uniform between 1 and $\min(\widetilde{\eta}_3^{(r)\prime}\widetilde{\eta}_3^{(r)}, \#(Z_{-x}^*) - \widetilde{\eta}_3^{(r)\prime}\widetilde{\eta}_3^{(r)})$, where $\#(Z_{-x}^*)$ denotes the number of elements in Z_{-x}^*. If $(\#(Z_{-x}^*) - \widetilde{\eta}_3^{(r)\prime}\widetilde{\eta}_3^{(r)}) = 0$ (which implies there are currently no potential instruments excluded from the model), we fix the candidate model equal to the the current one. Otherwise, among those elements of $\widetilde{\eta}_3^{(r)}$ that are currently one, $\hbar$ of them are randomly selected and changed to 0 (and the corresponding element in $\widetilde{\eta}_2^{(r)}$ will also be changed to 0). Similarly, among those elements of $\widetilde{\eta}_3^{(r)}$ that are currently zero, $\hbar$ of them are randomly selected and changed to 1 (and the corresponding element in $\widetilde{\eta}_2^{(r)}$ will also be changed to 1).

When $C = 2$ we will move potential instruments that are in z_i to x_i and viceversa. As in the case $C = 1$, we consider two types of movements: one that changes the number of instruments by just one, and another that changes the set of instruments while keeping the number of instruments the same. Let $\widehat{\eta}_3^{(r)}$ be those elements of $\eta_3^{(r)}$ whose corresponding element in $\eta_2^{(r)}$ is one (that is, $\widehat{\eta}_3^{(r)}$ corresponds to potential instruments that are currently in the model, either in x_i or in z_i). Conditional on $C = 2$ we will propose changes only to $\widehat{\eta}_3^{(r)}$, while keeping $\eta_2^{(r)}$ the same (that is, we are just moving potential instruments from z_i to x_i and viceversa). With probability ç we propose a move that changes the number of instruments: simply choose one element in $\widehat{\eta}_3^{(r)}$ randomly and change it. With probability (1-ç) we change a random number of elements in $\widehat{\eta}_3^{(r)}$ while leaving the value $(\eta_3^{(r)\prime}\eta_3^{(r)})$ constant. The number of elements to be changed, denote it as $\hbar$, is drawn from a uniform between 1 and $\min(\eta_3^{(r)\prime}\eta_3^{(r)}, \eta_2^{(r)\prime}\eta_2^{(r)} - \eta_3^{(r)\prime}\eta_3^{(r)})$, where $\#(Z^*)$ denotes the number of elements in Z^*. If $(\eta_2^{(r)\prime}\eta_2^{(r)} - \eta_3^{(r)\prime}\eta_3^{(r)} = 0)$ we fix $M^{(*)} = M^{(r)}$. Otherwise, among those elements of $\widehat{\eta}_3^{(r)}$ that are currently one, $\hbar$ of them are randomly selected and changed to 0. Similarly, among those elements of $\widehat{\eta}_3^{(r)}$ that are currently zero, $\hbar$ of them are randomly selected and changed to 1.

Thus, for each value of C $(1, 2, 3, 4)$ the proposal density we consider is symmetric and so it cancels out from the acceptance probability. Note that the proposal density might propose a new model $M^{(*)}$ such that $\eta_3^{(*)\prime}\eta_3^{(*)} < m$ (so the number of instruments in z_i is not enough for identification). By making the prior probability for such models equal to zero we make sure that such proposed models are always rejected.

Specification of prior in Drèze (1976)

Define $Y = (y_1, ..., y_N)'$, $Y_1 = (y_{11}, ..., y_{1N})'$, $Y_2 = (y_{21}, ..., y_{2N})'$, $X = (x_1, ..., x_N)'$, $Z = (z_1, ..., z_N)'$ and the cross-product matrices:

$$\begin{aligned} A_{YY} &= Y'Y \qquad A_{YX} = Y'X \qquad A_{YZ} = Y'Z \\ A_{XX} &= X'X \qquad A_{XZ} = X'Z \qquad A_{ZZ} = Z'Z \end{aligned}$$

Over-identified models with no restrictions on σ_{21}

With regard to Σ, it is tempting to use an improper non-informative prior for it. If there were no models with restrictions on the variance-covariance matrix we could use the non-informative prior: $p(\Sigma) \propto |\Sigma|^{-(m+1)/2}$, which

implies $p(\Omega) \propto |\Omega|^{-(m+1)/2}$. However, since the model space includes models with exogeneity restrictions we need to specify a proper prior for the relevant covariance parameters. Using the decomposition of Ω in (2), let us define:

$$\begin{aligned} \omega_{11\cdot 2} &= var(v_{1i}|v_{2i}) = \omega_{11} - \omega_{12}\Omega_{22}^{-1}\omega_{21} \\ \widetilde{\omega}_{21} &= \Omega_{22}^{-1}\omega_{21} \end{aligned} \tag{8}$$

There is a one-to-one mapping from Ω to $(\widetilde{\omega}_{21}, \Omega_{22}, \omega_{11\cdot 2})$ (e.g. Bauwens, Lubrano and Richard, 1999, p.305) and so we can fix the following prior specification on $(\widetilde{\omega}_{21}, \Omega_{22}, \omega_{11\cdot 2})$:

$$\begin{aligned} \widetilde{\omega}_{21} &\sim N(0, \underline{g}^e \omega_{11\cdot 2} I_m) \\ \Omega_{22} &\sim IW_m(\underline{S}_{22}, \underline{v}_{22}) \\ p(\omega_{11\cdot 2}) &\propto |\omega_{11\cdot 2}|^{-1} \end{aligned} \tag{9}$$

where $IW_m(\underline{S}_{22}, \underline{v}_{22})$ represents the inverted Wishart distribution with degrees of freedom equal to $\underline{v}_{22}$ and parameter matrix $\underline{S}_{22}$ (Bauwens et al. (1999 p. 305)). Let $\gamma_{\widetilde{E}}$ be a $d_{\widetilde{E}} \times 1$ vector containing the non-zero elements of γ. Following the parameterization in Drèze (1976) we specify a normal prior on $(\gamma'_{\widetilde{E}}, vec(\Pi_x)', vec(\Pi_{2z})')'$ such that $vec(\Pi_x)|\Omega \sim N(0, \underline{g}\underline{V}_{\Pi_x} \otimes \Omega)$, $\gamma_{\widetilde{E}}|\Omega \sim N(0, \underline{g}\omega_{11\cdot 2}\underline{A})$, $vec(\Pi_{2z})|\Omega \sim N(0, \underline{g}\underline{D} \otimes \Omega_{22})$, where $(\underline{g}, \underline{g}^e, \underline{V}_{\Pi_x}, \underline{A}, \underline{D})$ are prior parameters. It can be shown that $\Sigma_{22} = \Omega_{22}$, $\sigma_{11\cdot 2} = \omega_{11\cdot 2}$ and that $vec\begin{pmatrix} \beta' \\ \Pi_{2x} \end{pmatrix}|\Sigma \sim N(0, \underline{g}\underline{V}_{\Pi_x} \otimes \Sigma)$. The same type of prior can be used when there are restrictions on β (a zero restriction on β implies that the corresponding variable becomes an instrument or that it completely drops out from the system). In our empirical application of Section 5.2 we fixed: $\underline{g} = \underline{g}^e = N^2$, $\underline{V}_{\Pi_x} = A_{XX}^{-1}$, $\underline{A} = I_{d_{\widetilde{E}}}$, $\underline{D} = A_{ZZ}^{-1}$, $\underline{S}_{22} = \underline{g}^{-1} I_m$, $\underline{v}_{22} = m + 1$. In the analysis over the restricted model space in Section 5.1 we used the same prior except for $p(\Omega) \propto |\Omega|^{-(m+1)/2}$ and[17] $\underline{g} = N$. An advantage of this prior specification is that there are many analytical results for marginal posteriors. The following proposition summarizes results regarding marginal posterior densities that we use in our algorithm.

[17] We did not fix $\underline{g} = N^2$ in Section 5.1 because that would imply that the model that includes all potential instruments in z_i would get probability almost equal to 1. Conversely, we did not use $\underline{g} = N$ in Section 5.2 because with that choice Bayesian and frequentist estimates in the all-encompassing model would differ substantially.

Proposition 1 *Define S as:*

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where $S_{11} = (Y_1 - \widetilde{Z}\gamma)' M_x (Y_1 - \widetilde{Z}\gamma)$, $S_{12} = (Y_1 - \widetilde{Z}\gamma)' M_x (Y_2 - \widetilde{Z})$, $S_{21} = S_{12}'$ $S_{22} = (Y_2 - \widetilde{Z})' M_x (Y_2 - \widetilde{Z})$, $\widetilde{Z} = Z\Pi_{2z}'$ *and* $M_x = I_N - X\overline{V}_{\Pi_x} X'$, *with* $\overline{V}_{\Pi_x}$ *being defined below. Define also* $\overset{\bullet}{Y}_2$ *as the columns of* Y_2 *that correspond to the non-zero elements of* γ. *Similarly define* $\overset{\bullet}{\widetilde{Z}}$ *as the columns of* $\widetilde{Z}$ *that correspond to the non-zero elements of* γ. *Then, using the prior defined above, we can get the following posterior densities:*

$$vec(\Pi_x)|(\Omega, \gamma, \Pi_{2z}) \sim N(B_{\Pi_x}, \overline{V}_{\Pi_x} \otimes \Omega)$$

$$\widetilde{\omega}_{21}|(\Omega_{22}, \omega_{11\cdot 2}, \gamma, \Pi_{2z}) \sim N(B_{\widetilde{\omega}_{21}}, V_{\widetilde{\omega}_{21}}) \qquad \Omega_{22}|(\omega_{11\cdot 2}, \gamma, \Pi_{2z}) \sim IW_m(\overline{S}_{22}, \overline{v}_{22})$$

$$\omega_{11\cdot 2}|(\gamma, \Pi_{2z}) \sim IW_1(\overline{S}_{11\cdot 2}, \overline{v}_{11}) \qquad \gamma_{\widetilde{E}}|\Pi_{2z} \sim Mt_{d_{\widetilde{E}} \times 1}(M_\gamma, P_\gamma, Q_\gamma, v_\gamma)$$

where $Mt_{d_{\widetilde{E}} \times 1}(.)$ *refers to the multivariate Student distribution of dimension* $d_{\widetilde{E}} \times 1$ *(Bauwens, Lubrano and Richard (1999, p. 307)), and:*

$$\overline{V}_{\Pi_x} = \left(\left(\underline{g}\underline{V}_{\Pi_x}\right)^{-1} + A_{XX}\right)^{-1} \qquad V_{\widetilde{\omega}_{21}} = \omega_{11\cdot 2}\left(S_{22} + \frac{1}{\underline{g}^e} I_m\right)^{-1}$$

$$B_{\Pi_x} = vec((A_{YX} - \Pi_z A_{XZ}')\overline{V}_{\Pi_x}) \qquad B_{\widetilde{\omega}_{21}} = \left(S_{22} + \frac{1}{\underline{g}^e} I_m\right)^{-1} S_{21}$$

$$\overline{S}_{22} = S_{22} + \underline{S}_{22} + \underline{g}^{-1}\Pi_{2z}\underline{D}^{-1}\Pi_{2z}' \qquad \overline{v}_{22} = \underline{v}_{22} + k_2 + N$$

$$\overline{S}_{11\cdot 2} = S_{11} - S_{12}\left(S_{22} + \frac{1}{\underline{g}^e} I_m\right)^{-1} S_{21} + \underline{g}^{-1}\gamma_{\widetilde{E}}'\underline{A}^{-1}\gamma_{\widetilde{E}}$$

$$\overline{v}_{11} = N + d_{\widetilde{E}} \qquad v_\gamma = N$$

$$P_\gamma = \overset{\bullet}{\widetilde{Z}}{}' M_x \overset{\bullet}{\widetilde{Z}} + \frac{1}{\underline{g}}\underline{A}^{-1} - \overset{\bullet}{\widetilde{Z}}{}' M_x (Y_2 - \widetilde{Z})(\widehat{S}_{22})^{-1}(Y_2 - \widetilde{Z})' M_x \overset{\bullet}{\widetilde{Z}}$$

$$M_\gamma = P_\gamma^{-1}\left[\overset{\bullet}{\widetilde{Z}}{}' M_x Y_1 - \overset{\bullet}{\widetilde{Z}}{}' M_x (Y_2 - \widetilde{Z})(\widehat{S}_{22})^{-1}(Y_2 - \widetilde{Z})' M_x Y_1\right]$$

$$Q_\gamma = \left[Y_1' M_x Y_1 - Y_1' M_x (Y_2 - \widetilde{Z})(\widehat{S}_{22})^{-1}(Y_2 - \widetilde{Z})' M_x Y_1\right] - M_\gamma' P_\gamma M_\gamma$$

$$\widehat{S}_{22} = S_{22} + \frac{1}{\underline{g}^e} I_m$$

The posterior density conditional on Σ is:

$$\begin{pmatrix} \gamma_{\widetilde{E}} \\ vec(\Pi'_{2z}) \end{pmatrix} |\Sigma \sim N((\underline{T}+\overline{T})^{-1}(\underline{U}+\overline{U}), (\underline{T}+\overline{T})^{-1})$$

$$\overline{T} = \begin{pmatrix} a_{11} \otimes \overset{\bullet}{Y}'_2 M_x \overset{\bullet}{Y}_2 & a_{12} \otimes \overset{\bullet}{Y}'_2 M_x Z \\ a_{21} \otimes Z' M_x \overset{\bullet}{Y}_2 & A_{22} \otimes Z' M_x Z \end{pmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & A_{22} \end{pmatrix} \qquad a_{11} : 1 \times 1$$

$$\underline{T} = \begin{pmatrix} (\sigma_{11\cdot 2})^{-1} \left[\left(\underline{g}\underline{A}\right)^{-1} + \left(\underline{g}_e I_{d_{\widetilde{E}}}\right)^{-1} \right] & 0 \\ 0 & \Sigma_{22}^{-1} \otimes \underline{g}^{-1} \underline{D}^{-1} \end{pmatrix}$$

$$\underline{U} = \begin{pmatrix} -(\underline{g}_e \sigma_{11\cdot 2})^{-1} \overset{\bullet}{\widetilde{\sigma}}_{21} \\ 0_{k_2 m \times 1} \end{pmatrix} \qquad \overline{U} = \begin{pmatrix} vec(\overset{\bullet}{Y}'_2 M_x Y_2 a_{21}) + vec(\overset{\bullet}{Y}'_2 M_x Y_1 a_{11}) \\ vec(Z' M_x Y_1 a_{12}) + vec(Z' M_x Y_2 A_{22}) \end{pmatrix}$$

where $\widetilde{\sigma}_{21} = (\Sigma_{22})^{-1} \sigma_{21}$ and $\overset{\bullet}{\widetilde{\sigma}}_{21}$ contains only the rows of $\widetilde{\sigma}_{21}$ corresponding to the rows of γ where the non zero elements are located.

For $d_{\widetilde{E}} > 0$, $p(Y, \Pi_{2z})$ is given by:

$$\left|\underline{g}\underline{V}_{\Pi_x}\right|^{\frac{-(m+1)}{2}} \left|\overline{V}_{\Pi_x}\right|^{\frac{m+1}{2}} C_{IW}(\overline{S}_{22}, \overline{v}_{22}; m) C_{IW}(1, \overline{v}_{11}; 1) \left|S_{22} + \underline{g}_e^{-1} I_m\right|^{-1/2} \left|\underline{g}_e I_m\right|^{-1/2} \times$$
$$[C_{IW}(\underline{S}_{22}, \underline{v}_{22}; m)]^{-1} C_{M_t}(P_\gamma, Q_\gamma, v_\gamma; d_{\widetilde{E}}, 1) \times$$
$$\left|\underline{g}\underline{A}\right|^{-1/2} \left|\underline{g}\underline{D}\right|^{-m/2} |2\pi|^{-d_{\widetilde{E}}/2} |2\pi|^{-(k_2 m)/2} |2\pi|^{-N(m+1)/2}$$

where ($C_{IW}(.)$, $C_{M_t}(.)$) refers to the integrating constants of an Inverted Wishart and Matrix Student distribution respectively, as defined in Bauwens, Lubrano and Richard (1999, p. 305 and p. 307). For $d_{\widetilde{E}} = 0$ (i.e. all the elements of γ are restricted to be zero) the expression for $p(Y, \Pi_{2z})$ is the same but we need to write $\left|\overline{S}_{11\cdot 2}\right|^{-v_\gamma/2}$ instead of $C_{M_t}(P_\gamma, Q_\gamma, v_\gamma; d_{\widetilde{E}}, 1)$. Finally, $p(Y, \Sigma)$ is given by:

$$\left|\underline{g}\underline{V}_{\Pi_x}\right|^{\frac{-(m+1)}{2}} \left|\overline{V}_{\Pi_x}\right|^{\frac{m+1}{2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-1} Y' M_x Y)\right) \exp\left(\frac{1}{2}(\underline{U}+\overline{U})'(\underline{T}+\overline{T})^{-1}(\underline{U}+\overline{U})\right) \times$$
$$\left|(\underline{T}+\overline{T})^{-1}\right|^{1/2} \left|\underline{g}\sigma_{11\cdot 2}\underline{A}\right|^{-1/2} \left|\underline{g}\underline{D} \otimes \Sigma_{22}\right|^{-1/2} |2\pi|^{-N(m+1)/2} |\Sigma|^{-N/2} (\sigma_{11\cdot 2})^{-1} \times$$
$$[C_{IW}(\underline{S}_{22}, \underline{v}_{22}; m)]^{-1} |\Sigma_{22}|^{-(\underline{v}_{22}+m+1)/2} \exp(tr(\Sigma_{22}^{-1} \underline{S}_{22})) \times$$
$$\left|\underline{g}_e \sigma_{11\cdot 2} I_m\right|^{-1/2} \exp\left(-\frac{1}{2}\left(\underline{g}_e \sigma_{11\cdot 2}\right)^{-1} \widetilde{\sigma}'_{21} \widetilde{\sigma}_{21}\right)$$

Over-identified models with restrictions on σ_{12}

In the extreme case that $cov(v_{2i}, u_{1i}) = \sigma_{12} = 0$ (i.e. all variables are weakly exogenous) and, thus, $\widetilde{\sigma}_{12} = 0$, the prior for the remaining parameters in the model is exactly the same as above. In the intermediate case in which only some elements of σ_{12} are restricted, decompose y_{2i} into the weakly exogenous variables (y_{Xi}) and the exogenous variables (y_{Ei}). Furthermore, decompose y_{Xi} into $(y_{\widetilde{X}i}, y_{\overline{X}i})$, such that $y_{\overline{X}i}$ are those variables of y_{Xi} whose coefficient in the equation for y_{1i} is restricted to be zero. Then we can rewrite the model by including $y_{\overline{X}i}$ in z_i (as an instrument) and $y_{\widetilde{X}i}$ in x_i (as an exogenous regressor). This will result in a system of equations for y_{Ei} that is equivalent to (1) albeit of a lower dimension. We use the same prior outlined above, with the exception that the parts of ($\underline{D}$,$\underline{V}_{\Pi_x}$) that corresponds to $(y_{\overline{X}i}, y_{\widetilde{X}i})$ are chosen to be equal to the identity matrix. The system is completed with reduced form equations for y_{Xi} which depend on the original set of exogenous variables and with error terms that are independent from the error terms in the equations of (y_{1i}, y_{Ei}). The priors for the parameters in the equations for y_{Xi} are natural-conjugate[18] such that the marginall likelihood for this part of the system is known analytically (Zellner, 1971).

Specification of Cold Model

Using the prior of Drèze (1976) outlined above, the integrating constant of the conditional posterior $p(\gamma, \beta, \Pi_{2x}, \Sigma|Y, \Pi_{2z})$ can be calculated analytically. However, the integrating constant of $p(\Pi_{2z}|Y)$ and consequently the marginal likelihood $\pi(Y)$ are unknown. The cold model that we use has the same distribution for $p(\gamma, \beta, \Pi_{2x}, \Sigma|Y, \Pi_{2z})$, but uses an approximation for $p(\Pi_{2z}|Y)$ and $p(Y)$ that we denote as $p^c(\Pi_{2z}|Y)$ and $p^c(Y)$, where the super-index ^{c} denotes cold. Because Π_{2z} is asymptotically normal[19], we choose $p^c(\Pi_{2z}|Y)$ to be a multivariate Student density centered at the value of Π_{2z} that maximizes the posterior density $\widehat{\Pi}_{2z}$ (obtained using the methods outlined in Johansen (1988)) and with covariance matrix $P_\Pi \otimes \widehat{\Sigma}_{22}$, where $\widehat{\Sigma}_{22}$ is the value of Σ_{22}

[18] Specifically, conditional on the covariance matrix, mean coefficients follow a normal g-prior. The prior for the covariance matrix follows an inverted Wishart.

[19] Note that because Π_{2z} is a reduced form parameter, it is always identified, and hence the normal asymptotic approximation does not suffer from the problem of weak instruments.

that maximizes the posterior[20] and $P_{\Pi} = \left(\underline{g}^{-1}\underline{D}^{-1} + Z'M_xZ\right)^{-1}$. To see how this approximation of $p(\Pi_{2z}|Y)$ gives us also an approximation for $p(Y)$ first define:

$$p(Y, \Pi_{2z}) = \int p\left(\gamma, \beta, \Pi_{2x}, \Sigma, \Pi_{2z}\right) p\left(Y|\gamma, \beta, \Pi_{2x}, \Sigma, \Pi_{2z}\right) d(\gamma, \beta, \Pi_{2x}, \Sigma)$$

which can be obtained analytically when the prior of Drèze (1976) is used. Then note that:

$$p\left(\Pi_{2z}|Y\right) = \frac{p(Y, \Pi_{2z})}{p\left(Y\right)}$$

which implies that $p^c\left(Y\right)$ can be obtained as the ratio $\left(p(Y, \Pi_{2z})/p^c\left(\Pi_{2z}|Y\right)\right)$ evaluated at $\Pi_{2z} = \widehat{\Pi}_{2z}$. In order to design the RJMCMC algorithm, we need to know the joint density of parameters and data in the cold model, and this is defined as:

$$p^c(\gamma, \beta, \Pi_{2x}, \Sigma, \Pi_{2z}, Y) = p^c\left(Y\right) p(\gamma, \beta, \Pi_{2x}, \Sigma|Y, \Pi_{2z}) p^c(\Pi_{2z}|Y)$$

Prior specification in the Strachan and Inder's (2004) approach

Over-identified models with no restrictions on σ_{12}

Because this prior was originally proposed for the Vector Error Correction Model, we give details here of how it can be adapted to the incomplete simultaneous equations model. Decompose y_{2i} as $(y_{\widetilde{E}i}, y_{\overline{E}i})$, where $y_{\widetilde{E}i}$ are the variables that enter into the equation for y_{1i} with a non-zero coefficient, and $y_{\overline{E}i}$ are those whose coefficients are restricted to be zero. Similarly, decompose the error term v_{2i} into $(v_{\widetilde{E}i}, v_{\overline{E}i})$. Referring to the notation used in (1) let the rows of Π_{2x} that correspond to $(y_{\widetilde{E}i}, y_{\overline{E}i})$ be denoted as $(\Pi_{\widetilde{E}x}, \Pi_{\overline{E}x})$, respectively. Similarly, decompose the rows of Π_{2z} into $(\Pi_{\widetilde{E}z}, \Pi_{\overline{E}z})$. With this notation let us rewrite (1) as:

$$\begin{aligned} y_{1i} &= \gamma'_{\widetilde{E}} y_{\widetilde{E}i} + \beta' x_i + u_{1i} \\ y_{\widetilde{E}i} &= \Pi_{\widetilde{E}x} x_i + \Pi_{\widetilde{E}z} z_i + v_{\widetilde{E}i} \\ y_{\overline{E}i} &= \Pi_{\overline{E}x} x_i + \Pi_{\overline{E}z} z_i + v_{\overline{E}i} \end{aligned} \tag{10}$$

[20] To be more precise, $(\widehat{\Pi}_{2z}, \widehat{\Sigma}_{22})$ do not maximize the posterior density, but maximize the product of the likelihood times the priors of Π_x and Π_{2z} only. In models with restrictions on γ, these are ignored at the time of maximizing the posterior.

The reduced form can be written as:

$$\begin{pmatrix} y_{1i} \\ y_{\widetilde{E}i} \end{pmatrix} = \Pi_x^* x_i + \Pi_z^* z_i + v_{1i}^* \qquad (11)$$
$$y_{\overline{E}i} = \Pi_{\overline{E}x} x_i + \Pi_{\overline{E}z} z_i + v_{\overline{E}i}$$

where:

$$\Pi_x^* = \begin{pmatrix} \pi_{1x} \\ \Pi_{\widetilde{E}x} \end{pmatrix} = \begin{pmatrix} \beta' + \gamma'_{\widetilde{E}} \Pi^*_{\widetilde{E}x} \\ \Pi^*_{\widetilde{E}x} \end{pmatrix} \qquad \Pi_z^* = \begin{pmatrix} \gamma'_{\widetilde{E}} \Pi_{\widetilde{E}z} \\ \Pi_{\widetilde{E}z} \end{pmatrix} = \begin{pmatrix} \gamma'_{\widetilde{E}} \\ I_{d_{\widetilde{E}}} \end{pmatrix} \Pi_{\widetilde{E}z}$$
$$v_{1i}^* = \begin{pmatrix} u_{1i} + \gamma'_{\widetilde{E}} v_{\widetilde{E}i} \\ v_{\widetilde{E}i} \end{pmatrix}$$
$$\Omega^* = E\left(\begin{pmatrix} v_{1i}^* \\ v_{\overline{E}i} \end{pmatrix} \begin{pmatrix} v_{1i}^* & v'_{\overline{E}i} \end{pmatrix} \right) = \begin{pmatrix} \Omega_{11}^* & \Omega_{1\overline{E}}^* \\ \Omega_{\overline{E}1}^* & \Omega_{\overline{E}\overline{E}}^* \end{pmatrix}$$

Note that the matrix that is subject to rank restriction is Π_z^*. Following Koop, Leon-Gonzalez and Strachan (2010) let us introduce a non-identified matrix α_2 of dimension $d_{\widetilde{E}} \times d_{\widetilde{E}}$, where $d_{\widetilde{E}}$ is the dimension of $y_{\widetilde{E}i}$, and rewrite Π_z^* as:

$$\Pi_z^* = \begin{pmatrix} \gamma'_{\widetilde{E}} \\ I_{d_{\widetilde{E}}} \end{pmatrix} \Pi^*_{\widetilde{E}z} = \begin{pmatrix} \gamma'_{\widetilde{E}} \\ I_{d_{\widetilde{E}}} \end{pmatrix} \alpha_2 \alpha_2^{-1} \Pi^*_{\widetilde{E}z} = \begin{pmatrix} \gamma'_{\widetilde{E}} \alpha_2 \\ \alpha_2 \end{pmatrix} \alpha_2^{-1} \Pi^*_{\widetilde{E}z} = \alpha \beta'$$
$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \gamma'_{\widetilde{E}} \alpha_2 \\ \alpha_2 \end{pmatrix} : (d_{\widetilde{E}} + 1) \times d_{\widetilde{E}} \qquad \beta = \Pi^{*\prime}_{\widetilde{E}z} \left(\alpha_2^{-1}\right)' : k_{2j} \times d_{\widetilde{E}}$$

Thus, for a given value of α_2 there is a one-to-one mapping between the parameters in (1) and the parameters in (11). Therefore, it is possible to derive some of the properties that this prior implies on the structural parameters of (1). In particular, conditional on Ω, the implied prior for γ is a type of Cauchy with no prior moments. In this way the prior is quite non-informative, but still proper. The implied prior for $\Pi^*_{\widetilde{E}z}$ is a multivariate version of the variance-gamma distribution analyzed by Madan.and Seneta (1990). This distribution gives more weight to the tails and center of the distribution, at the expense of the middle range. Using standard rules for Jacobians (e.g. Muirhead (1982, p.57)) it can be verified that the Jacobian J from $(\alpha_2, \gamma, \beta, \Pi_{2x}, \Sigma, \Pi_{2z})$ to $(\alpha, \beta, \Pi_x^*, \Pi_{\overline{E}x}, \Pi_{\overline{E}z}, \Omega^*)$ is:

$$J = \left|\alpha_2 \alpha_2'\right|^{-\frac{1}{2}(k_{2j}-1)}$$

We proceed to put priors directly on the parameters of (11):

$$vec(\alpha)|\Omega^* \sim N\left(0, I_{d_{\widetilde{E}}} \otimes \Omega^*_{11}\right) \qquad vec(\beta') \sim N\left(0, \underline{gD} \otimes I_{d_{\widetilde{E}}}\right)$$

$$vec\begin{pmatrix} \Pi^*_x \\ \Pi_{\overline{E}x} \end{pmatrix}|\Omega^* \sim N\left(0, \underline{gV}_{\Pi_x} \otimes \Omega^*\right)$$

$$vec(\Pi_{\overline{E}z})|\Omega^* \sim N\left(0, \underline{gD} \otimes \Omega^*_{\overline{EE}}\right)$$

We fix the prior parameters ($\underline{g}$,$\underline{D}$,$\underline{V}_{\Pi_x}$) in the same way as we did in the prior of Drèze. The prior for Ω^* is also as in (9). An advantage of this prior specification is that it is possible to draw directly from the conditional posteriors. In particular, the conditional posteriors $(vec(\alpha)', vec(\Pi^*_x)', vec(\Pi_{\overline{E}x})'$ $, vec(\Pi_{\overline{E}z})')'|\Omega^*$ and $(vec(\beta')',\ vec(\Pi^*_x)', vec(\Pi_{\overline{E}x})'\ , vec(\Pi_{\overline{E}z})')'|\Omega^*$ are both normal, while $\Omega^*|(\alpha, \beta, \Pi^*_x,\ \Pi_{\overline{E}x}, \Pi_{\overline{E}z})$ is an inverted Wishart (see Koop, Leon-Gonzalez and Strachan (2010) for details).

Over-identified models with restrictions on σ_{12}

We follow the same strategy as we did with the prior of Drèze (1976). Using the same notation, y_{2i} was decomposed as $(y_{\widetilde{X}i}, y_{\widetilde{E}i}, y_{\overline{E}i}, y_{\overline{X}i})$. We can rewrite the model by including $y_{\overline{X}i}$ in z_i (i.e. as an instrument) and $y_{\widetilde{X}i}$ in x_i (i.e. as an exogenous regressor). This will result in a system of equations for $(y_{1i}, y_{\widetilde{E}i}, y_{\overline{E}i})$ that is equivalent to (11) albeit of a smaller dimension. Therefore we use the same prior for the parameters for this smaller system of equations as in the case of no restrictions on σ, with the exception that the parts of ($\underline{D}$,$\underline{V}_{\Pi_x}$) that corresponds to $(y_{\overline{X}i}, y_{\widetilde{X}i})$ are chosen to be equal to the identity matrix. The system is completed with reduced form equations for y_{Xi} which depend on the original set of exogenous variables and with error terms that are independent from the error terms in the equations of (y_{1i}, y_{Ei}). As discussed above, the priors for the parameters in the equations for y_{Xi} are natural-conjugate such that the marginall likelihood for this part of the system is known analytically (Zellner, 1971).

Prior specification in the approach of Kleibergen and van Dijk (1998)

We use the prior in expression (3.13) of Kleibergen and van Dijk (1998). Using the notation in their paper, the prior parameters that we need to choose are (P, A, h, G). In Section 5.1 we have fixed them as:

$$P = 0, \qquad A = \underline{g}^{-1}\begin{pmatrix} A_{XX} & A_{XZ} \\ A'_{XZ} & A_{ZZ} \end{pmatrix}, \qquad h = 5, \qquad G = 0.01 I_{m+1} \qquad \underline{g} = N$$

We use the numerical methods in Kleibergen and Paap (2002, Section 5) to calculate the Bayes factors. However, part of the Bayes factor calculation involves calculating c_r, which is the normalizing constant of the prior. As explained by Kleibergen and Paap (2002, Section 5), this could be either calculated using draws from the prior or in the case of an improper prior it could be set equal to $c_r = (2\pi)^{-1/2(k_2-m)^2}$. Even though our prior is proper, for simplicity we fix c_r equal to $(2\pi)^{-1/2(k_2-m)^2}$.