Journal of Econometrics 171 (2012) 237–250

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian model averaging in the instrumental variable regression model[☆]

Gary Koop[a], Roberto Leon-Gonzalez[b,*], Rodney Strachan[c]

[a] University of Strathclyde, United Kingdom
[b] National Graduate Institute for Policy Studies, Japan
[c] The Australian National University, Australia

ARTICLE INFO

Article history:
Available online 7 July 2012

JEL classification:
C11
C30

Keywords:
Bayesian
Endogeneity
Simultaneous equations
Reversible jump Markov chain Monte Carlo

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 such as exogeneity or over-identification. We illustrate our methods in a returns-to-schooling application.

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, 1983; Sargan, 1958). Given the large number of possible models, the repeated application of diagnostic tests will result in similar distorted size

[☆] All authors are fellows of the Rimini Centre for Economic Analysis and Strachan is also a Fellow of The Centre for Applied Macroeconomic Analysis. The authors would like to thank Frank Kleibergen and other participants 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.
E-mail address: rlg@grips.ac.jp (R. Leon-Gonzalez).

[1] See, among many others, Fernandez et al. (2001) and the references cited therein.

[2] Two related papers are Cohen-Cole et al. (2009) and Lenkoski et al. (forthcoming) 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. Tobias and Li (2004) does BMA in a returns to schooling example similar to the one we use, but this paper does not address endogeneity concerns.

0304-4076/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2012.06.005

properties as arise in the regression model with exogenous regressors. Since estimates of structural parameters 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.

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 et al., 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.[3] 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 develop is very flexible and can be easily adapted to handle any of the popular approaches to Bayesian inference in IV models such as Drèze (1976), Kleibergen and Van Dijk (1998) and Strachan and Inder (2004). We describe in detail how the algorithm works in the context of two popular Bayesian approaches to instrumental variables and reduced rank regression. These are the classic approach of Drèze (1976) and the modern approach of Strachan and Inder (2004).[4] We also show how, if desired, the RJMCMC algorithm can be easily coded to produce results for several different 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 applies our methods to a returns-to-schooling example based on Card (1995) and Section 5 concludes.

2. Modelling choices in the incomplete simultaneous equations model

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

$$y_{1i} = \gamma' y_{2i} + \beta' x_i + u_{1i} \qquad (1)$$
$$y_{2i} = \Pi_{2x} x_i + \Pi_{2z} z_i + v_{2i}$$

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

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

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

$$y_i = \Pi_x x_i + \Pi_z z_i + v_i \qquad (2)$$

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

$$\Pi_x = \begin{pmatrix} \pi_{1x} \\ \Pi_{2x} \end{pmatrix} = \begin{pmatrix} \gamma' \Pi_{2x} + \beta' \\ \Pi_{2x} \end{pmatrix}$$

$$\Pi_z = \begin{pmatrix} \pi_{1z} \\ \Pi_{2z} \end{pmatrix} = \begin{pmatrix} \gamma' \\ I_m \end{pmatrix} \Pi_{2z}$$

$$\Sigma = E\left(\begin{pmatrix} u_{1i} \\ v_{2i} \end{pmatrix} \begin{pmatrix} u_{1i} & v_{2i}' \end{pmatrix}\right) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$\Omega = E(v_i v_i') = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \Omega_{22} \end{bmatrix} = \begin{pmatrix} 1 & \gamma' \\ 0 & I_m \end{pmatrix} \Sigma \begin{pmatrix} 1 & 0 \\ \gamma & I_m \end{pmatrix}$$

$$\Pi_x : (m+1) \times k_{1j} \qquad \Pi_z : (m+1) \times k_{2j}.$$

The subindex j stands for the jth 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

$$\pi_{1x} - \gamma' \Pi_{2x} = \beta' \quad \text{and} \qquad (3)$$
$$\pi_{1z} - \gamma' \Pi_{2z} = 0. \qquad (4)$$

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 more equations than we need to identify γ and so the equation is over-identified. If $k_{2j} < m$ so rank $(\Pi_z) < m$, there are fewer equations than necessary to identify γ and thus 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 $k_{2j} > m$ and $(k_{2j} - m_j)$ columns of the $m \times k_{2j}$ matrix Π_{2z} are zero, or, if rank $(\Pi_{2z}) = m_j < m$, 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 $m \times m_j$ and ϱ is $m_j \times k_{2j}$ 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 fewer 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:

[3] To avoid confusion, note that some of the related literature uses different terminology where the space of distributions to be simulated from is augmented with hot distributions, while the actual target distribution is the distribution with the lowest temperature. See for example Kou et al. (2006) and Hoogerheide et al. (2011).

[4] 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 et al. (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 et al. (2010). The working paper version of this paper, available on Gary Koop's website, provides full details of how Kleibergen and Van Dijk (1998)'s prior can be used with our algorithm.

- 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 $x_i \in Z^* \cup X^*$).

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 (and this is what we do in our empirical work). 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. Or 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 using a large model space based on the classic returns-to-schooling paper of Card (1995) and associated data set. Details and acronyms 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 vector y_{2i} includes education (*ED*76), experience squared (*EXPER*2) and a measure of ability (*KWW*).[5]

Table 1
Posterior probabilities and 2SLS estimates for the return to schooling for 16 models with restrictions on $(\gamma_E, \gamma_X, \sigma_E, \sigma_X)$. 0 indicates that the corresponding parameter is restricted to be zero, while 1 indicates that it is unrestricted. $T = 0, 1, 2$ corresponds to the 3 alternative priors described in the Appendix. The column SDDR uses the Savage-Dickey-Density-Ratio to calculate the probabilities, assuming the prior $T = 1$ in the unrestricted model and a proper prior for $\sigma_{11\cdot 2}$. The 2SLS standard error of the return to schooling is inside brackets. See Appendix for other technical details.

γ_E	γ_X	σ_E	σ_X	$T = 0$	$T = 1$	$T = 2$	SDDR	2SLS
0	0	0	0	0.00	0.00	0.00	0.00	0.04 (0.003)
1	0	0	0	0.00	0.00	0.00	0.00	0.07 (0.005)
0	1	0	0	0.00	0.00	0.00	0.24	0.08 (0.005)
1	1	0	0	0.00	0.00	0.00	0.02	0.08 (0.005)
0	0	1	0	0.00	0.00	0.00	0.00	0.04 (0.003)
1	0	1	0	0.04	0.04	0.98	0.01	0.08 (0.005)
0	1	1	0	0.53	0.53	0.02	0.09	0.08 (0.005)
1	1	1	0	0.00	0.00	0.00	0.00	0.08 (0.005)
0	0	0	1	0.00	0.00	0.00	0.00	0.04 (0.003)
1	0	0	1	0.02	0.02	0.00	0.03	0.07 (0.005)
0	1	0	1	0.41	0.41	0.00	0.37	0.07 (0.005)
1	1	0	1	0.00	0.00	0.00	0.01	0.07 (0.005)
0	0	1	1	0.00	0.00	0.00	0.06	0.04 (0.003)
1	0	1	1	0.00	0.00	0.00	0.08	0.15 (0.042)
0	1	1	1	0.00	0.00	0.00	0.08	0.08 (0.014)
1	1	1	1	0.00	0.00	0.00	0.01	0.13 (0.048)

Including just-identified models, our model space involves[6] C_j^{13} for $j = 3, \ldots, 13$ 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

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

models if we ignore exogeneity restrictions and restrictions on γ. But there are 2^m of each of these resulting in $64N^A$ models. Adding all these models together yields approximately 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.

To motivate the use of BMA abstracting first from the difficulties posed by a large model space, let us initially consider a reduced model space of just 16 models. As in the well-known specification of Card (1995), we exclude *KWW* and *IQ* from the analysis and use only 4 instruments: *NEARC*2, *NEARC*4, *NEARC*4*A* and age squared (*AGE*762). The regressors in X^* are those in Table A.1 in the Data Appendix and the 16 models are defined by considering only zero restrictions on $\gamma = (\gamma_E, \gamma_X)$ (i.e. coefficients of *ED*76 and *EXPER*2) and $\sigma_{12} = (\sigma_E, \sigma_X)$ (i.e. exogeneity restrictions). Table 1 shows posterior model probabilities calculated under alternative priors for the parameters and the 2SLS estimate of the return to schooling under each specification. All approaches indicate that models with

[5] Card (1995) also includes experience in y_{2i}, and age (*AGE*76) as an instrument, with experience defined as: (age – education – 6). However, this specification is not suitable for Bayesian analysis, because it implies a singular covariance matrix for v_i. Instead, we include age as a regressor (i.e. in X^*) and exclude experience from the analysis (but still include *EXPER*2 in y_2). Our specification is just a reparameterization of that of Card (1995), and in our case the return to schooling, which is the key structural parameter of interest, is given by the sum of the coefficients of *ED*76 and *AGE*76.

[6] C_c^b denotes "b choose c": the number of sets of c elements chosen without replacement from a set of b elements. Alternative notation for this function is $\binom{b}{c}$.

more than one restriction are much more probable, and favour models where the estimated return to schooling is lower and more precisely estimated than in the unrestricted model (estimate of 0.08 versus 0.13, with standard error 0.005 versus 0.048). Even though the model ($\sigma_E = \sigma_X = 0$) can be rejected, there is substantial uncertainty regarding the exogeneity restrictions. The posterior probability of ($\sigma_E = 0$) ranges between 0 and 0.66 while that of ($\sigma_X = 0$) ranges between 0.36 and 1. This illustrates that BMA helps in finding restrictions that are empirically relevant and in handling situations in which there is more than one model with non-negligible posterior probability.

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 develop an RJMCMC algorithm to sample from the joint posterior defined over the parameter and the model spaces.

RJMCMC algorithms take draws of models and parameters where the models may have different dimensions. This is achieved by matching the dimensions of the parameters in the current model and the proposed model at the proposal step. That is, the vector of parameters, θ, of the current model is augmented with another random vector of parameters, u, to produce a vector (θ, u). Further, the vector of parameters, θ^*, of the proposed model is augmented with a vector of parameters, u^*, to produce a vector (θ^*, u^*). A crucial condition is that the dimensions of (θ, u) and (θ^*, u^*) are the same. The acceptance probability for a move then requires a one to one mapping between (θ, u) and (θ^*, u^*) and densities for u and u^* conditional upon θ and θ^* be specified.

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 $j = 1, \ldots, N^{\text{mod}}$ models and each model depends on parameters Π_j which determine the conditional mean of the incomplete simultaneous equations model (i.e. $\Pi_j = (\beta', \gamma', \text{vec}(\Pi_{2x})', \text{vec}(\Pi_{2z})')'$) 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 $r = 1, \ldots, R$. 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 et al. (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 $p(\Pi, M|Y, \Sigma)$ and $p(\Sigma|Y, \Pi, M)$. To sample from $p(\Pi, M|Y, \Sigma)$ 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 $(\Pi^{(*)}, M^{(*)})$ with an appropriate acceptance probability. If accepted, we have $(\Pi^{(r)}, M^{(r)}) = (\Pi^{(*)}, M^{(*)})$. If not, then $(\Pi^{(r)}, M^{(r)}) = (\Pi^{(r-1)}, M^{(r-1)})$.

For the SUR model, it can be shown that choosing $q(\Pi^*|\Sigma, M^*) = p(\Pi^*|Y, \Sigma, M^*)$ leads to the most efficient RJMCMC algorithm. As we have seen, since HDM use a normal prior for Π, $p(\Pi^*|Y, \Sigma, M^*)$ 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.[7]

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^*

[7] For more details on RJMCMC algorithms, see for example Green (1995) or Waagepetersen and Sorensen (2001).

(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 overcome 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, 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.[8] 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, are 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).[9] We will not explain these approaches here, nor motivate their advantages over Drèze (1976). However, it is worth noting that sensitivity to the prior can be an important issue in any Bayesian analysis and in models such as the IV model, where identification (or weak identification) is a concern, this issue can be of particular importance. The priors of Kleibergen and Van Dijk (1998) and Strachan and Inder (2004) are two recent examples of priors which treat identification issues in a more satisfactory way than earlier approaches. The reader is referred to these papers for an explanation of these points. Furthermore, there may be other priors that the researcher may wish to work with. In this section we describe an MCMC algorithm which should work with any such prior. This is useful for the researcher who wishes to work with a prior other than the one of Drèze (1976) or who wishes to do a prior sensitivity analysis over a range of prior choices. In the Technical Appendix, we provide complete details of how this approach would work for the prior of Strachan and Inder (2004) and this prior is used in our empirical work. However, we stress that this algorithm will work with a much wider range of priors.

Let $p^* (\Pi, \Sigma)$ be a prior which is different from the prior used in Drèze (1976). The latter 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. Application: 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,

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

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

Table 2
Frequentist and Bayesian estimates and 95% confidence intervals of returns to schooling in the all encompassing model.

OLS	0.05	0.06	0.07
2SLS	0.03	0.11	0.18
LIML	0.04	0.15	0.26
$T = 0$	0.05	0.12	0.21
$T = 1$	0.03	0.11	0.21
$T = 2$	−0.08	0.14	0.52

Table 3
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.

	$I_e = 0$			$I_e = 1$		
$T = 0$	0.00	0.01	0.11	0.01	0.10	0.12
$T = 1$	0.00	0.01	0.11	0.01	0.10	0.12
$T = 2$	0.00	0.01	0.02	0.01	0.02	0.11
$T = 0, I_r = 1$	0.07	0.11	0.13	0.09	0.10	0.12
$T = 1, I_r = 1$	0.07	0.11	0.13	0.09	0.11	0.12
$T = 2, I_r = 1$	0.02	0.11	0.13	0.09	0.10	0.12
$T = 0, I_c = 1$	0.00	0.01	0.08	0.09	0.11	0.12
$T = 1, I_c = 1$	0.00	0.01	0.08	0.09	0.11	0.12
$T = 2, I_c = 1$	0.00	0.01	0.02	0.09	0.11	0.12
$T = 0, I_r = 1, I_c = 1$	0.07	0.11	0.13	0.09	0.11	0.12
$T = 1, I_r = 1, I_c = 1$	0.07	0.11	0.13	0.09	0.11	0.12
$T = 2, I_r = 1, I_c = 1$	0.02	0.11	0.13	0.09	0.11	0.12
$T = 0, I_b = 1$	0.07	0.08	0.09	0.04	0.04	0.05
$T = 1, I_b = 1$	0.07	0.08	0.09	0.04	0.04	0.05
$T = 2, I_b = 1$	0.07	0.08	0.09	0.04	0.18	0.35
$T = 0, I_b = 1, I_r = 1$	0.07	0.08	0.09	0.09	0.18	0.33
$T = 1, I_b = 1, I_r = 1$	0.07	0.08	0.09	0.09	0.18	0.38
$T = 2, I_b = 1, I_r = 1$	0.07	0.08	0.09	0.09	0.18	0.32
$T = 0, I_b = 1, I_c = 1$	0.06	0.08	0.09	0.04	0.04	0.05
$T = 1, I_b = 1, I_c = 1$	0.06	0.08	0.09	0.04	0.04	0.05
$T = 2, I_b = 1, I_c = 1$	0.07	0.08	0.09	0.04	0.04	0.24
$T = 0, I_b = 1, I_r = 1, I_c = 1$	0.06	0.08	0.09	0.05	0.16	0.33
$T = 1, I_b = 1, I_r = 1, I_c = 1$	0.06	0.08	0.09	0.05	0.16	0.36
$T = 2, I_b = 1, I_r = 1, I_c = 1$	0.07	0.08	0.09	0.05	0.15	0.29

our model space for this application will include approximately 10^{16} models. We consider all models a priori equally likely. The Technical Appendix describes the prior on the parameters of each model. Recall the definitions from Table 1 that $T = 0, 1, 2$ indicates the three model temperatures and also corresponds to three Bayesian estimation procedures: an approximation, one using the prior of Drèze (1976) and one using the prior of Strachan and Inder (2004), respectively. We run the algorithm for 535 000 iterations, after a burn-in of 70 000. To check the convergence of the RJMCMC algorithm we calculate the Total Visited Probability (George and McCulloch, 1997). This is an estimate of the joint posterior probability of the set of models visited by the algorithm and in our case was over 0.99 for all $T = 0, 1, 2$.[10]

We begin by presenting results which are not based on Bayesian model averaging. Table 2 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 wider credible intervals with $T = 2$ are due to properties of this prior noted by Strachan and Inder (2004) and, in particular, 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_b, I_e, I_r, 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_b takes value 1 when y_{2i} excludes KWW and when Z^* (the set of potential instruments) includes only: $NEARC2$, $NEARC4$, $NEARC4A$ and $AGE762$ (the basic specification of Card, 1995). 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_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 3. Results for BMA over the full model space are given in the rows labelled $T = 0, 1, 2$ in the columns under $I_e = 0$. For all of our three temperatures, our point estimate of the returns to schooling is 0.01 and 95% credible intervals are fairly narrow. This point estimate is substantively lower than those in Table 2. For instance, with LIML we found a point estimate of 0.15, and 0.01 is outside the LIML 95% confidence interval.

The other estimates in Table 3, 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 2. For example, the posterior median for $T = 0, 1$ is 10%. For $T = 2$ the 95% credible interval is the same as for $T = 0, 1$, but the posterior median is lower (2%) and thus nearer to the case in which exogeneity restrictions are allowed.

As we found when we used the reduced set of 16 models (Table 1), we now find that conditional on $I_b = 1$ the probability that education is endogenous is at least 0.58 (Table 4). However, when $I_b = 0$ we are effectively introducing a measure of ability (KWW) as a control variable, and the probability of education being endogenous decreases to almost 0. This is consistent with the findings of Blackburn and Neumark (1995, p. 228), who find little evidence that education is endogenous once they controlled for test scores. Fig. 1, which shows BMA posterior densities (conditional on $I_e = 1, I_b = 0$) of the correlations between v_2 and u_1, supports this view.

Tables 5 and 6 show the probability that each variable enters each category and 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 5 and 6 indicate that some are included as instruments, but most are excluded altogether from the model.

[10] In order to calculate the total visited probability one needs to define a large set of models A that encompass those visited by the algorithm (B). We defined A as a set of models that form a large neighborhood (in the sense described in the Appendix) of the best model in B. The marginal likelihood is calculated for all the models in A and the total visited probability is the joint posterior probability of B over that of A.

[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 (2005) test fails to reject that 2SLS estimates might be subject to 30% or more bias due to weak instruments. The Stock–Yogo test is a test of the null hypothesis that a set of instruments is weak. This test uses the Cragg and Donald (1993) test statistic, although we note that the Cragg–Donald test is a test of the null hypothesis of underidentification. The definition of weak instruments for the Stock–Yogo test depends upon the relative bias in the 2SLS versus the OLS estimators.

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

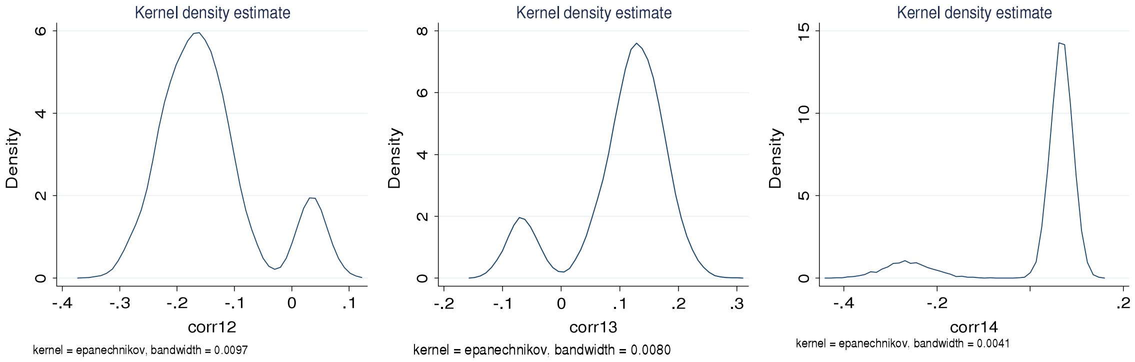

Fig. 1. Posterior density for the correlation between u and v conditional on $I_e = 1$ and $T = 1$. From left to right the correlations correspond to u and the error terms of ED76, EXPER2 and KWW.

Table 4
Posterior probabilities of ED76, EXPER2 and KWW being endogenous.

	$I_b = 0$			$I_b = 1$		
	$T = 0$	$T = 1$	$T = 2$	$T = 0$	$T = 1$	$T = 2$
ED76	0.05	0.05	0.00	0.58	0.58	0.99
EXPER2	0.10	0.09	0.05	0.43	0.43	0.01
KWW	0.96	0.96	1.00	–	–	–

Table 5
Posterior probability of variables in Z^* entering in the model as an instrument (in z).

	$I_b = 0$			$I_b = 1$		
	$T = 0$	$T = 1$	$T = 2$	$T = 0$	$T = 1$	$T = 2$
EDFDUM1	0.50	0.50	0.54	0.12	0.12	0.60
EDFDUM2	0.34	0.34	0.31	0.67	0.67	0.28
EDFDUM3	0.00	0.00	0.00	0.00	0.00	0.01
EDFDUM4	0.01	0.01	0.01	0.01	0.01	0.01
EDFDUM5	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM6	0.05	0.05	0.04	0.09	0.09	0.05
EDFDUM7	0.00	0.00	0.00	0.00	0.00	0.00
EDFDUM8	0.01	0.01	0.01	0.00	0.00	0.01
NEARC4	0.01	0.01	0.00	0.01	0.01	0.00
NEARC2	0.00	0.00	0.00	0.00	0.00	0.00
NEARC4A	0.08	0.08	0.09	0.10	0.10	0.04
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

Similarly, BMA allows the variables in X^* to be either exogenous regressors or be excluded from the model. Table 6 indicates most are excluded from the model. The 2SLS estimate of the returns to schooling in the best model selected by the RJMCMC is 0.013, with 95% confidence interval being (0.005, 0.022).[13]

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 2. 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 1.4% for each of the 3 temperatures (Table 3). 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.[14] A further difference between BMA and non-BMA

Table 6
Posterior probability of being included as a regressor in the first structural equation (in x or y_2).

	$I_b = 0$			$I_b = 1$		
	$T = 0$	$T = 1$	$T = 2$	$T = 0$	$T = 1$	$T = 2$
ED76	0.04	0.04	0.00	0.60	0.59	0.07
EXPER2	0.08	0.11	0.00	0.30	0.30	0.00
KWW	0.89	0.86	1.00	0.10	0.11	0.93
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

[13] The best model, which coincides under the 3 temperatures, has posterior probability 0.45, 0.43 and 0.52 under priors $T = 0, 1, 2$, respectively. The best 10 models have a joint posterior probability of 93%, 94% and 98% for $T = 0, 1, 2$, respectively.

[14] The diagnostics of STATA indicated that instruments in the best models are strong (e.g. Stock–Yogo test). Furthermore, the marginal likelihood of the best model is at least 60 points larger, on the log scale, than that of the best non-identified model, indicating strong identification. By best non-identified model we mean the best SUR model. We found this model by running the HMD algorithm described in Section 3.1. The best non-identified model is characterized by using only two variables from Z^* (*IQ*, *AGE*762) and 4 from X^* (*AGE*76, *BLACK*, *SMSA*76*R*, *DADED*).

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.

5. 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.

Appendix A. 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 44 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[a]	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^*
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^*

[a] 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.

Appendix B. Technical appendix

B.1. 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:

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

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:

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

where

$$\begin{aligned} p(Y, \theta_{1M}|M) &= \int p(Y, \theta_{1M}, \theta_{2M}|M) d\theta_{2M} \\ &= \int p(\theta_{1M}, \theta_{2M}|M) p(Y|\theta_{1M}, \theta_{2M}, M) d\theta_{2M}. \end{aligned}$$

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 $\Pi = (\beta', \gamma', \text{vec}(\Pi_{2x})', \text{vec}(\Pi_{2z})')'$). 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 rth 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 $\tau_1^* \le \tau_2^*$. 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.[15]

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 $u \ge \tau_1^*$: (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. 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 $\tau_1^* \le u \le \tau_2^*$: (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 $q(M|M^{(r)}, T^{(r+1)} = 1)$. 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 $\tau_2^* \le u$: (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 $p(\Pi^{(r+1)}, \Sigma^{(r+1)}|Y, M^{(r+1)}, T^{(r+1)} = 2)$.

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

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 $p(\cdot|T = 2)$ using the same parameterization as $p(\cdot|T = 1)$. 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 $p(\cdot|T = 2)$ 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 $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_E)$ 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 η_3 is one, then the corresponding element in η_2 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 5 types of movements: $(C = 1)$ Change only η_1, $(C = 2)$ Change only (η_2, η_3), $(C = 3)$ Change only η_3, $(C = 4)$ Change only η_4 and $(C = 5)$ Change only (η_4, η_E). Under $T = 0$, $q(M^*|M, T = 0)$ is such that C can take values (1, 2, 3, 5) each with equal probability (1/4). When $T = 1$, C takes values (1, 2, 3, 4), each with probability 1/4. Conditional on $C = 4$, one of the elements in $\eta_4^{(r)}$ 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). Conditional on $C = 5$, with probability ς_5 we randomly select one element of the concatenated vector $(\eta_4', \eta_E')'$ and change it. With probability $(1 - \varsigma_5)$ we randomly select two elements and change them. Conditional on $C = 1$, with probability ς_1 one random element of η_1 is selected and changed. With probability $(1-\varsigma_1)$ we change a random number (maximum 4) of randomly selected elements.[16]

When $C = 2$ we will propose movements that take instruments (i.e. variables that belong to Z^*) in and out of the model. That is, 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. Variables that belong to Z^* but are currently in x_i will not be affected. Let $\widetilde{\eta}_2^{(r)}$ be those elements of $\eta_2^{(r)}$ that correspond to potential instruments that are currently not in x_i. With the intention of improving convergence speed, when $C = 2$ we do not only consider increasing or decreasing the number of instruments 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 ς_2 one of the elements in $\widetilde{\eta}_2^{(r)}$ is chosen randomly and its current value is changed. This is a move that changes the number of instruments. When an element of $\widetilde{\eta}_2^{(r)}$ is chosen, the corresponding value in $\eta_2^{(r)}$ *and* in $\eta_3^{(r)}$ will be changed. 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 $(1-\varsigma_2)$ 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 $(1-\varsigma_2)$ 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 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 $C = 3$ we will move potential instruments that are in z_i to x_i and vice versa. As in the case $C = 2$, 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 $C = 3$ 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 vice versa). With probability ς_3 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 $(1 - \varsigma_3)$ 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 $\min(\eta_3^{(r)\prime}\eta_3^{(r)}, \eta_2^{(r)\prime}\eta_2^{(r)} - \eta_3^{(r)\prime}\eta_3^{(r)})$. 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, 5) 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.

B.2. *Specification of prior in Drèze (1976)*

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

$$A_{YY} = Y'Y \qquad A_{YX} = Y'X \qquad A_{YZ} = Y'Z$$
$$A_{XX} = X'X \qquad A_{XZ} = X'Z \qquad A_{ZZ} = Z'Z.$$

[16] We do this by choosing first a random number (maximum 2) of elements that are currently one and change them to zero. Then we choose again a random number (maximum 2) of elements that are currently zero and change them to one. Note that the number of ones changed could be different from the number of zeros changed.

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:

$$\omega_{11\cdot 2} = \mathrm{var}(\upsilon_{1i}|\upsilon_{2i}) = \omega_{11} - \omega_{12}\Omega_{22}^{-1}\omega_{21}$$

$$\widetilde{\omega}_{21} = \Omega_{22}^{-1}\omega_{21}. \tag{8}$$

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

$$\widetilde{\omega}_{21} \sim N(0, \underline{g}^{e}\omega_{11\cdot 2}I_m) \tag{9}$$

$$\Omega_{22} \sim \mathrm{IW}_m(\underline{S}_{22}, \underline{\upsilon}_{22})$$

$$p(\omega_{11\cdot 2}) \propto |\omega_{11\cdot 2}|^{-1}$$

where $\mathrm{IW}_m(\underline{S}_{22}, \underline{\upsilon}_{22})$ represents the inverted Wishart distribution with degrees of freedom equal to $\underline{\upsilon}_{22}$ 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 $(\gamma_{\widetilde{E}}', \mathrm{vec}(\Pi_x)', \mathrm{vec}(\Pi_{2z})')'$ such that $\mathrm{vec}\,(\Pi_x)\,|\Omega \sim N(0, \underline{g}\underline{V}_{\Pi_x} \otimes \Omega)$, $\gamma_{\widetilde{E}}|\Omega \sim N(0, \underline{g}\omega_{11\cdot 2}\underline{A})$, $\mathrm{vec}(\Pi_{2z})|\Omega \sim N(0, \underline{g}\underline{D}\otimes\Omega_{22})$, where $(\underline{g}, \underline{g}^{e}, \underline{V}_{\Pi_x}, \underline{A}, \underline{D})$ are prior hyper-parameters. It can be shown that $\Sigma_{22} = \Omega_{22}, \sigma_{11\cdot 2} = \omega_{11\cdot 2}$ and that $\mathrm{vec}\begin{pmatrix}\beta' \\ \Pi_{2x}\end{pmatrix}|\Sigma \sim N(0, \underline{g}\underline{V}_{\Pi_x} \otimes \Sigma)$. 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 applications, we set $\underline{g} = \underline{g}^{e}$ and estimated this hyper-parameter by choosing the value with the highest marginal likelihood from the set N, N^2 and N^3. In both our restricted model and full model space applications this led to $\underline{g} = \underline{g}^{e} = N^2$. In our application we set $\underline{V}_{\Pi_x} = A_{XX}^{-1}, \underline{A} = I_{d_{\widetilde{E}}}, \underline{D} = A_{ZZ}^{-1}, \underline{S}_{22} = \underline{g}^{-1}I_m,\ \underline{\upsilon}_{22} = m + 1$. An advantage of these prior specifications 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.

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* $\dot{Y}_2$ *as the columns of* Y_2 *that correspond to the non-zero elements of* γ. *Similarly define* $\dot{\widetilde{Z}}$ *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:*

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

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

$$\Omega_{22}|(\omega_{11\cdot 2}, \gamma, \Pi_{2z}) \sim \mathrm{IW}_m(\overline{S}_{22}, \overline{\upsilon}_{22})$$

$$\omega_{11\cdot 2}|(\gamma, \Pi_{2z}) \sim \mathrm{IW}_1(\overline{S}_{11\cdot 2}, \overline{\upsilon}_{11})$$

$$\gamma_{\widetilde{E}}|\Pi_{2z} \sim Mt_{d_{\widetilde{E}}\times 1}(M_\gamma, P_\gamma, Q_\gamma, \upsilon_\gamma)$$

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

$$\overline{V}_{\Pi_x} = \left(\left(\underline{g}\underline{V}_{\Pi_x}\right)^{-1} + A_{XX}\right)^{-1}$$

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

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

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

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

$$\overline{\upsilon}_{11} = N + d_{\widetilde{E}} \qquad \upsilon_\gamma = N$$

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

$$M_\gamma = P_\gamma^{-1}\left[\dot{\widetilde{Z}}'M_xY_1 - \dot{\widetilde{Z}}'M_x(Y_2 - \widetilde{Z})(\widehat{S}_{22})^{-1}(Y_2 - \widetilde{Z})'M_xY_1\right]$$

$$Q_\gamma = \left[Y_1'M_xY_1 - Y_1'M_x(Y_2 - \widetilde{Z})(\widehat{S}_{22})^{-1}(Y_2 - \widetilde{Z})'M_xY_1\right] - M_\gamma'P_\gamma M_\gamma$$

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

The posterior density conditional on Σ *is:*

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

$$\overline{T} = \begin{pmatrix} a_{11} \otimes \dot{Y}_2'M_x\dot{Y}_2 & a_{12} \otimes \dot{Y}_2'M_xZ \\ a_{21} \otimes Z'M_x\dot{Y}_2 & A_{22} \otimes Z'M_xZ \end{pmatrix}$$

$$\Sigma^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & A_{22} \end{pmatrix} \qquad a_{11} : 1 \times 1$$

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

$$\underline{U} = \begin{pmatrix} -(\underline{g}_e\sigma_{11\cdot 2})^{-1}\dot{\widetilde{\sigma}}_{21} \\ 0_{k_2m\times 1} \end{pmatrix}$$

$$\overline{U} = \begin{pmatrix} \mathrm{vec}(\dot{Y}_2'M_xY_2a_{21}) + \mathrm{vec}(\dot{Y}_2'M_xY_1a_{11}) \\ \mathrm{vec}(Z'M_xY_1a_{12}) + \mathrm{vec}(Z'M_xY_2A_{22}) \end{pmatrix}$$

where $\widetilde{\sigma}_{21} = (\Sigma_{22})^{-1}\sigma_{21}$ *and* $\dot{\widetilde{\sigma}}_{21}$ *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:

$$\left|\underline{g}\underline{V}_{\Pi_x}\right|^{\frac{-(m+1)}{2}}\left|\overline{V}_{\Pi_x}\right|^{\frac{m+1}{2}} C_{\mathrm{IW}}(\overline{S}_{22}, \overline{\upsilon}_{22}; m)C_{\mathrm{IW}}(1, \overline{\upsilon}_{11}; 1)$$
$$\times\left|S_{22} + \underline{g}_e^{-1}I_m\right|^{-1/2}\left|\underline{g}_e I_m\right|^{-1/2}$$
$$\times\left[C_{\mathrm{IW}}(\underline{S}_{22}, \underline{\upsilon}_{22}; m)\right]^{-1} C_{M_t}(P_\gamma, Q_\gamma, \upsilon_\gamma; d_{\widetilde{E}}, 1)$$
$$\times\left|\underline{g}\underline{A}\right|^{-1/2}\left|\underline{g}\underline{D}\right|^{-m/2}|2\pi|^{-d_{\widetilde{E}}/2}|2\pi|^{-(k_2m)/2}|2\pi|^{-N(m+1)/2}$$

where $(C_{\mathrm{IW}}(\cdot), C_{M_t}(\cdot))$ refers to the integrating constants of an Inverted Wishart and Matrix Student distribution respectively, as defined in Bauwens et al. (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 $\left|\overline{S}_{11\cdot 2}\right|^{-\upsilon_\gamma/2}$ instead of $C_{M_t}(P_\gamma, Q_\gamma, \upsilon_\gamma; d_{\widetilde{E}}, 1)$. Finally, $p(Y, \Sigma)$ is given by:

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

Identified models with restrictions on σ_{12}.

In the extreme case that $\mathrm{cov}(\upsilon_{2i}, u_{1i}) = \sigma_{12} = 0$ (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 endogenous 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 correspond 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 of 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[17] such that the marginal likelihood for this part of the system is known analytically (Zellner, 1971).

B.3. 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 $p(Y)$ 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. We approximate the marginal posterior of Π_{2z} by choosing[18] $p^c(\Pi_{2z}|Y)$ to be a multivariate Student density with N degrees of freedom 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} that maximizes the posterior[19] 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:

$$\begin{aligned}
p(Y, \Pi_{2z}) = \int & p\,(\gamma, \beta, \Pi_{2x}, \Sigma, \Pi_{2z}) \\
& \times p\,(Y|\gamma, \beta, \Pi_{2x}, \Sigma, \Pi_{2z})\, d(\gamma, \beta, \Pi_{2x}, \Sigma)
\end{aligned}$$

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

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

which implies that $p^c\,(Y)$ can be obtained as the ratio $(p(Y, \Pi_{2z})/p^c(\Pi_{2z}|Y))$ 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:

$$\begin{aligned}
& p^c(\gamma, \beta, \Pi_{2x}, \Sigma, \Pi_{2z}, Y) \\
&\quad = p^c\,(Y)\, p(\gamma, \beta, \Pi_{2x}, \Sigma|Y, \Pi_{2z}) p^c(\Pi_{2z}|Y).
\end{aligned}$$

B.4. Prior specification in the Strachan and Inder's (2004) approach

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 υ_{2i} into $(\upsilon_{\widetilde{E}i}, \upsilon_{\overline{E}i})$. 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:

$$\begin{aligned}
y_{1i} &= \gamma_{\widetilde{E}}' y_{\widetilde{E}i} + \beta' x_i + u_{1i} \qquad (10) \\
y_{\widetilde{E}i} &= \Pi_{\widetilde{E}x} x_i + \Pi_{\widetilde{E}z} z_i + \upsilon_{\widetilde{E}i} \\
y_{\overline{E}i} &= \Pi_{\overline{E}x} x_i + \Pi_{\overline{E}z} z_i + \upsilon_{\overline{E}i}.
\end{aligned}$$

The reduced form can be written as:

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

where:

$$\Pi_x^* = \begin{pmatrix} \pi_{1x} \\ \Pi_{\widetilde{E}x} \end{pmatrix} = \begin{pmatrix} \beta' + \gamma_{\widetilde{E}}'\Pi_{\widetilde{E}x} \\ \Pi_{\widetilde{E}x} \end{pmatrix}$$

$$\Pi_z^* = \begin{pmatrix} \gamma_{\widetilde{E}}'\Pi_{\widetilde{E}z} \\ \Pi_{\widetilde{E}z} \end{pmatrix} = \begin{pmatrix} \gamma_{\widetilde{E}}' \\ I_{d_{\widetilde{E}}} \end{pmatrix} \Pi_{\widetilde{E}z}$$

$$\upsilon_{1i}^* = \begin{pmatrix} u_{1i} + \gamma_{\widetilde{E}}'\upsilon_{\widetilde{E}i} \\ \upsilon_{\widetilde{E}i} \end{pmatrix}$$

$$\Omega^* = E\left(\begin{pmatrix} \upsilon_{1i}^* \\ \upsilon_{\overline{E}i} \end{pmatrix} \begin{pmatrix} \upsilon_{1i}^* & \upsilon_{\overline{E}i}' \end{pmatrix}\right) = \begin{pmatrix} \Omega_{11}^* & \Omega_{1\overline{E}}^* \\ \Omega_{\overline{E}1}^* & \Omega_{\overline{EE}}^* \end{pmatrix}.$$

Note that the matrix that is subject to rank restriction is Π_z^*. Following Koop et al. (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:

$$\begin{aligned}
\Pi_z^* = \begin{pmatrix} \gamma_{\widetilde{E}}' \\ I_{d_{\widetilde{E}}} \end{pmatrix} \Pi_{\widetilde{E}z} &= \begin{pmatrix} \gamma_{\widetilde{E}}' \\ I_{d_{\widetilde{E}}} \end{pmatrix} \\
&\quad \times \alpha_2\alpha_2^{-1}\Pi_{\widetilde{E}z} = \begin{pmatrix} \gamma_{\widetilde{E}}'\alpha_2 \\ \alpha_2 \end{pmatrix} \alpha_2^{-1}\Pi_{\widetilde{E}z} = \alpha\widetilde{\beta}'
\end{aligned}$$

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \gamma_{\widetilde{E}}'\alpha_2 \\ \alpha_2 \end{pmatrix} : (d_{\widetilde{E}} + 1) \times d_{\widetilde{E}}$$

$$\widetilde{\beta} = \Pi_{\widetilde{E}z}'\left(\alpha_2^{-1}\right)' : k_{2j} \times d_{\widetilde{E}}.$$

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

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

[19] 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.

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

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

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

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

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 $(\alpha, \widetilde{\beta}, \Pi^*_x, \Pi_{\overline{E}x}, \Pi_{\overline{E}z}, \Omega^*)$ is:

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

We 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 $(\mathrm{vec}(\alpha)', \mathrm{vec}(\Pi^*_x)', \mathrm{vec}(\Pi_{\overline{E}x})', \mathrm{vec}(\Pi_{\overline{E}z})')'|\Omega^*$ and $(\mathrm{vec}(\widetilde{\beta}')', \mathrm{vec}(\Pi^*_x)', \mathrm{vec}(\Pi_{\overline{E}x})'\ , \mathrm{vec}(\Pi_{\overline{E}z})')'|\Omega^*$ are both normal, while $\Omega^*|(\alpha, \widetilde{\beta}, \Pi^*_x, \Pi_{\overline{E}x}, \Pi_{\overline{E}z})$ is an inverted Wishart (see Koop et al., 2010, for details).

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_{12}, with the exception that the parts of $(\underline{D}, \underline{V}_{\Pi_x})$ that correspond 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 of 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 marginal likelihood for this part of the system is known analytically (Zellner, 1971).

B.5. Other technical details

With respect to the calculations in Table 1, the SDDR assumes that the priors in models with restrictions are derived from the conditional prior in the unrestricted model given the restrictions. In our case the conditional prior given exogeneity restrictions is derived using the parameterization $(\widetilde{\sigma}_E, \widetilde{\sigma}_X)$, where $\widetilde{\sigma}_E = (\sigma_{EE\cdot X})^{-1}\sigma_E, \widetilde{\sigma}_X = (\sigma_{XX\cdot E})^{-1}\sigma_X$, $\sigma_{XX\cdot E}$ denotes the conditional variance of EXPER2 given ED76 and $\sigma_{EE\cdot X}$ is the conditional variance of ED76 given EXPER2. Hence the exogeneity restrictions become $\widetilde{\sigma}_E = 0$ and/or $\widetilde{\sigma}_X = 0$. For the SDDR calculations we assumed a proper prior for $\sigma_{11\cdot 2}$ with 2 degrees of freedom in the unrestricted model: $\sigma_{11\cdot 2} \sim \mathrm{IW}_1(N^{-2}, 2)$. Marginal likelihoods under $T = 0$ were calculated analytically, under $T = 1$ using the method of Chib (1995) and under $T = 2$ using importance sampling (using the posterior under $T = 1$ as importance density). The numerator of the SDDR is calculated by averaging a conditional posterior over posterior draws from a Gibbs sampler. The denominator used draws from the prior. The prior parameters $(\underline{g}, \underline{g}^e)$ were both fixed equal to N^2, because that gave higher marginal likelihoods than fixing them to N or N^3. The Gibbs sampler for $T = 1$ draws all parameters (except for Π_{2z}) conditional on Π_{2z}, and then draws all parameters (except for Σ) conditional on Σ.

With respect to the Total Visited Probability, for $T = 0$ the neighborhood A of the best model (see footnote 10) consisted of models that had any restriction on γ or σ_{12}, used just 3 instruments in z_i (whichever set of 3, with the other 10 excluded from the model) and whose set of regressors (x_i) differed from those in the best model by at most one. This made a total of 530 816 models. For $T = 1, 2$ we defined A to be the best 1000 models of these 530 816 models (i.e. 'best' in terms of $T = 0$ posterior model probabilities).

References

Anderson, T., 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., 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., Richard, J.-F., 1999. Bayesian Inference in Dynamic Econometric Models. Oxford University Press, Oxford.

Blackburn, M., Neumark, D., 1995. Are OLS estimates of the return to schooling biased downward? another look. Review of Economics and Statistics 77, 217–230.

Card, D., 1995. Using geographic variation in college proximity to estimate the return to schooling. In: Christofides, L., Grant, E., Swidinsky, R. (Eds.), Aspects of Labour Market Behaviour: Essays in Honour of John Vandekamp. University of Toronto Press, Toronto.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

Cohen-Cole, E., Durlauf, S., Fagan, J., Nagin, D., 2009. Model uncertainty and the deterrent effect of capital punishment. American Law and Economics Review 11, 335–369.

Cragg, J., Donald, S., 1993. Testing identifiability and specification in instrumental variable models. Econometric Theory 9, 222–240.

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

Drèze, J., Richard, J.-F., 1983. Bayesian analysis of simultaneous equations systems. In: Handbook of Econometrics, Vol. 1. North Holland, Amsterdam, pp. 517–598.

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

George, E., McCulloch, R., 1997. Approaches for Bayesian variable selection. Statistica Sinica 7, 339–373.

Geyer, C., 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 ed. Prentice-Hall, New Jersey.

Hausman, J., 1983. Specification and estimation of simultaneous equations models. In: Handbook of Econometrics, Vol. 1. North Holland, Amsterdam, pp. 391–448.

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

Holmes, C.C., Held, L., 2006. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analysis 1 (1), 145–168.

Hoogerheide, L., Opschoor, A., Van Dijk, H.K., 2011. A class of adaptive EM-based importance sampling algorithms for efficient and robust posterior and predictive simulation. Tinbergen Institute Discussion Paper TI 2011-004.

Johansen, S., 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12, 231–254.

Kleibergen, F., Van Dijk, H.K., 1998. Bayesian simultaneous equations analysis using reduced rank structures. Econometric Theory 14, 699–744.

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

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

Kou, S., Zhou, Q., Wong, W., 2006. Equi-energy sampler with applications in statistical inference and statistical mechanics. The Annals of Statistics 34, 1581–1619.

Lenkoski, A., Eicher, T.S., Raftery, A.E., 2011. Two-stage Bayesian model averaging in the endogenous variable model. Econometric Reviews (forthcoming).

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

Madan, D., Seneta, E., 1990. The variance-gamma (V-G) model for share market returns. Journal of Business 63, 511–524.

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

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

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

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

Stock, J., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Andrews, D., Stock, J. (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, Cambridge, pp. 80–108.

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

Tobias, J., Li, M., 2004. Returns to schooling and Bayesian model averaging: a union of two literatures. Journal of Economic Surveys 18, 153–180.

Waagepetersen, R., 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. Wiley, New York.

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