TI 2012-098/III
Tinbergen Institute Discussion Paper

Bayesian Analysis of Instrumental Variable Models: Acceptance-Rejection within Direct Monte Carlo

Arnold Zellner[a]
Tomohiro Ando[b]
Nalan Baştürk[c,d,f]
Lennart Hoogerheide[e,f]
Herman K. van Dijk[c,e,f]

[a] *(posthumous) Booth School of Business, University of Chicago, USA;*
[b] *Graduate School of Business Administration, Keio University, Japan;*
[c] *Econometric Institute, Erasmus University Rotterdam, The Netherlands;*
[d] *The Rimini Centre for Economic Analysis, Rimini, Italy;*
[e] *Department of Econometrics, VU University Amsterdam, The Netherlands;*
[f] *Tinbergen Institute, The Netherlands.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Bayesian Analysis of Instrumental Variable Models: Acceptance-Rejection within Direct Monte Carlo[1]

Arnold Zellner [a], Tomohiro Ando [b], Nalan Baştürk [c,d], Lennart Hoogerheide [e,f], and Herman K. van Dijk [c,e,f]

[a] *(posthumous) Booth School of Business, University of Chicago, USA*

[b] *Graduate School of Business Administration, Keio University, Japan*

[c] *Econometric Institute, Erasmus University Rotterdam, The Netherlands*

[d] *The Rimini Centre for Economic Analysis, Rimini, Italy*

[e] *Department of Econometrics, Vrije Universiteit Amsterdam, The Netherlands*

[f] *Tinbergen Institute, The Netherlands*

September 21, 2012

[1] This paper started through intense, lively discussions between Arnold Zellner and Herman K. van Dijk in April 2010 when the latter was visiting Chicago. We note that, given the untimely death of Arnold Zellner in August 2010, he has not been involved in the empirical illustrations and the revision of this paper. However, all co-authors feel that Arnold's influence on the topic of this paper has been so important that as a credit to his enormous positive activity on the topic we wish to maintain him as co-author although he is not responsible for any errors in the empirical analysis. The authors are indebted to the editors Essie Maasoumi and Ehsan Soofi and two anonymous referees for very helpful comments that greatly helped in the revision of an earlier version of this paper.

Abstract

We discuss Bayesian inferential procedures within the family of instrumental variables regression models and focus on two issues: existence conditions for posterior moments of the parameters of interest under a flat prior and the potential of Direct Monte Carlo (DMC) approaches for efficient evaluation of such possibly highly non-elliptical posteriors. We show that, for the general case of m endogenous variables under a flat prior, posterior moments of order r exist for the coefficients reflecting the endogenous regressors' effect on the dependent variable, if the number of instruments is greater than $m+r$, even though there is an issue of local non-identification that causes non-elliptical shapes of the posterior. This stresses the need for efficient Monte Carlo integration methods. We introduce an extension of DMC that incorporates an acceptance-rejection sampling step within DMC. This *Acceptance-Rejection within Direct Monte Carlo* (ARDMC) method has the attractive property that the generated random drawings are independent, which greatly helps the fast convergence of simulation results, and which facilitates the evaluation of the numerical accuracy. The speed of ARDMC can be easily further improved by making use of parallelized computation using multiple core machines or computer clusters. We note that ARDMC is an analogue to the well-known "Metropolis-Hastings within Gibbs" sampling in the sense that one 'more difficult' step is used within an 'easier' simulation method. We compare the ARDMC approach with the Gibbs sampler using simulated data and two empirical data sets, involving the settler mortality instrument of Acemoglu et al. (2001) and father's education's instrument used by Hoogerheide et al. (2012a). Even without making use of parallelized computation, an efficiency gain is observed both under strong and weak instruments, where the gain can be enormous in the latter case.

Key words: Instrumental variables, Bayesian inference, Direct Monte Carlo, Acceptance-Rejection, numerical standard errors
JEL Classification: C11, C15, C26, C36

1 Introduction

In many areas of economics and other sciences, models are specified that contain instantaneous feedback mechanisms between variables. An important example is the market system where prices and quantities are jointly determined. The Simultaneous Equations Model (SEM), that incorporates this mechanism, was systematically analyzed in the nineteen forties and early nineteen fifties and documented in the well-known Cowles Commission Monographs (Koopmans, 1950; Hood and Koopmans, 1950) and has been widely employed to analyze the behavior of markets, macroeconomic and other multivariate systems. Inference on a complete system of the SEM is rather involved and very sensitive to the assumptions, see e.g. Bauwens and Van Dijk (1990); Van Dijk (2003). Therefore, Zellner, Bauwens, and Van Dijk (1988) proceeded with a more tractable and robust analysis of a single equation of the SEM. This model can be linked to the so-called Instrumental Variable (IV) regression model,

where the issue of endogeneity, another expression for immediate feedback mechanisms, is extensively investigated (see e.g. Angrist and Krueger (1991)). A third basic econometric model is the Errors in Variables (EV) model where a measurement error in all variables is explicitly specified. The interesting feature of these three models, SEM, IV and EV models, is their common statistical structure, namely, a possible strong correlation between a right hand side variable in an equation and the disturbance of that equation. This creates, however, an important problem for Bayesian econometric inference compared to such inference in the basic regression model. We note that in case of the basic linear regression model using a flat prior, the coefficients have a Student-t posterior distribution. In the IV models that we investigate, the posterior densities of the parameters of interest are a product of a Student-t density and a polynomial or rational function. Then one faces two issues. First, do analytical properties of posterior distributions exist? Second, how can one efficiently evaluate posterior properties numerically by Monte Carlo methods, especially if the shape of the posterior may be highly non-elliptical? We emphasize that some of our results on these two issues carry over to the SEM and EV model but that for space considerations we restrict ourselves to the IV model. For more details on the similarity of the mathematical structure of the IV model, EV model and SEM we refer to Zellner et al. (2011).

The first issue on conditions for the existence of posterior moments relates to the well-known condition of non-singularity of the parameter matrix that reflects the effect of the instrumental variables on the possibly endogenous regressors, and to the number of instrumental variables compared to the number of endogenous regressors. We present an overview of the joint, conditional and marginal posterior distributions (and posterior moments) in the IV model with $m \geq 1$ possibly endogenous regressors under a flat prior. We show that in the case of over-identification, or more precisely in the presence of $m + r + 1$ instruments, posterior moments of order r exist for the coefficients that reflect the endogenous regressors' effect on the dependent variable, even though a parameter matrix may become singular. Further, for the coefficients that indicate the instruments' effect on the endogenous regressors the first few moments exist for any case of over-identification; to the best of our knowledge, an analysis of the posterior moments of these coefficients is novel. This is contrary to earlier suggestions in the literature stating that the posterior of an IV model with flat prior may be improper due to the unboundedness of the marginal posterior; see for instance Hoogerheide et al. (2007). In case of over-identification Gibbs sampling is feasible; the region of locally non-identified parameter values is not an absorbing state (if identified parameter values are used as initial values), contrary to a claim by Kleibergen and Van Dijk (1998).

Although the posterior is proper in case of a sufficient number of instruments, one faces in empirical econometrics many situations where the data information is weak in the sense of weak identifiability or weak instrumental variables, strong endogeneity and the lack of many available instruments. In these situations, the posterior may often have substantial mass near and/or at the boundary of the parameter region. Examples of data sets yielding such posterior shapes are given in section 4; see also De Pooter et al. (2008). The empirical issue is the following: given that much

data information may exist at or near the boundary of singularity, the researcher may not want to exclude this information by a strong informative prior that focuses on the center of the parameter space and seriously down-weights or truncates relevant information near the boundary. This situation does not only occur in weak instrument models but also in unit root and cointegration models where the issue of near-market efficiency is related to the occurrence of time series with near unit roots. One also faces this issue in factor models. In all these situations one may encounter a most important problem for empirical research, that is, the appearance of highly non-elliptical shapes of the posterior distributions.

Monte Carlo methods have been successfully applied for the computation of posterior and predictive results. Typically, one uses an *indirect* Monte Carlo method where one makes use of a correction mechanism like a rejection step, an importance weighting step or Markov Chain steps. For details on these methods we refer to standard textbooks like Geweke (2005). The obvious reason is that *direct* sampling – simulating independent drawings without a rejection step, an importance weighting step or Markov Chain steps – is typically not feasible. Very attractive properties of direct simulation are that it is straightforward to apply and that the generated random drawings are independent, which greatly helps the speed of convergence of simulation results, and facilitates the computation of accurate numerical standard errors or predictive likelihoods. Further, the computations can be easily performed in a parallelized fashion, which may yield another huge reduction of computing time on multiple core machines or computer clusters. In earlier work Zellner and Ando labeled this approach Direct Monte Carlo (DMC); see Zellner and Ando (2008), Zellner and Ando (2010a), Zellner and Ando (2010b) and Ando and Zellner (2010).

The important issue in the present paper is to determine whether the posterior distribution studied for IV models allows for DMC. Specifically, we discuss the applicability of DMC approaches in IV models with several possibly endogenous regressors, multiple instruments, and Gaussian errors under a flat prior. We emphasize that for models with multiple endogenous regressors complete direct sampling is not possible. We introduce an acceptance-rejection sampling step within the DMC method to simulate from a low-dimensional marginal posterior distribution of coefficients of interest. In order to obtain a suitable candidate distribution we use a novel adaptation of the *Mixture of t by Importance Sampling weighted Expectation Maximization* (MitISEM) method of Hoogerheide et al. (2012b). Until now the MitISEM procedure has only been used to construct an importance or candidate density for Importance Sampling or the independence chain Metropolis-Hastings algorithm. Our novel adaptation aims at a high acceptance rate in the acceptance-rejection method rather than a low variance of the Importance Sampling weights. Due to the flexibility of the MitISEM approach and the low dimension of the marginal posterior, we are able to achieve rather high acceptance rates, i.e., higher than 45%. We label our method *Acceptance-Rejection within Direct Monte Carlo* (ARDMC) and note that ARDMC is an analogue to the well-known "Metropolis-Hastings within Gibbs" sampling method in the sense that one 'more difficult' step is used within an 'easier' simulation method.

In order to evaluate the efficiency of ARDMC, we compare our approach with the Gibbs sampler using simulated data and two empirical data sets, involving the settler mortality instrument of Acemoglu et al. (2001) and father's education's instrument used by Hoogerheide et al. (2012a). Even without making use of parallelized computation, an efficiency gain is observed both under strong and weak instruments, where the gain can be enormous in the latter case. For illustrative purposes, we also present the posterior shapes.

The remainder of this paper is organized as follows. Section 2 considers the joint, conditional and marginal posterior distributions (and their moments) in the IV model with $m \geq 1$ possibly endogenous regressors. Section 3 discusses DMC and our proposed Acceptance-Rejection within Direct Monte Carlo (ARDMC) approach. Section 4 shows applications where the performance of ARDMC and Gibbs sampling is investigated. Section 5 discusses further possibilities of ARDMC, stressing the scope of the method. Section 6 concludes.

2 IV model with m possibly endogenous regressors under a flat prior: Existence of proper conditional and marginal posterior distributions and posterior moments

In this section, we present an analysis of the joint, conditional and marginal posterior distributions (and posterior moments) in the following IV model with $m \geq 1$ possibly endogenous regressors under a flat prior:

$$y_t = x_t\beta + u_t, \tag{1}$$
$$x_t = z_t\Pi + v_t, \tag{2}$$

for $t = 1, \ldots, T$, where y_t is the dependent variable, x_t is the $1 \times m$ vector of (possibly) endogenous explanatory variables, z_t is the $1 \times k$ vector of instruments; β $(m \times 1)$ and Π $(k \times m)$ contain model parameters; u_t (1×1) and v_t $(1 \times m)$ contain disturbances. Finally, $(u_t, v_t')' \sim NID(0_{(m+1)\times 1}, \Sigma)$ with $(m+1) \times (m+1)$ positive-definite symmetric matrix $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12}' & \Sigma_{22} \end{pmatrix}$, where σ_{11}, σ_{12} and Σ_{22} are 1×1, $1 \times m$ and $m \times m$ matrices, respectively. [1]

The matrix representation of the model in (1) and (2) is:

$$y = X\beta + u, \tag{3}$$
$$X = Z\Pi + V, \tag{4}$$

[1] The model (1)-(2) may include exogenous explanatory variables w_t $(1 \times n)$ in both equations. In that case, we assume a flat prior for the coefficients at w_t, and these coefficients are marginalized out of the posterior distribution using analytical integration. This amounts to replacing y_t, x_t and z_t by their residuals after regression on w_t, and replacing T by $T - n$.

where $y = (y_1, \ldots, y_T)'$, $X = (x_1', \ldots, x_T')'$, $Z = (z_1', \ldots, z_T')'$, $u = (u_1, \ldots, u_T)'$, $V = (v_1', \ldots, v_T')'$, and $(u', \mathrm{vec}(V)')' \sim N(0_{(T\times(m+1))\times 1}, \Sigma \otimes I_T)$. We assume that the data matrix $(y \;\; X \;\; Z)$ has full column rank $m + k + 1$. The posterior density under a flat prior $p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-h/2}$ with $h = m + 2$ is:

$$p(\beta, \Pi, \Sigma \mid y, X, Z) \propto |\Sigma|^{-(T+m+2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left((u \;\; V)'\,(u \;\; V)\,\Sigma^{-1}\right)\right\}, \qquad (5)$$

where $u = y - X\beta$ and $V = X - Z\Pi$. Highly non-elliptical posterior shapes may result from the local non-identification of β if Π does not have full column rank, which is easily seen from the restricted reduced form:

$$y = Z\Pi\beta + \tilde{u}, \qquad (6)$$
$$X = Z\Pi + V, \qquad (7)$$

with $\tilde{u} \equiv V\beta + u$, where β drops out from (6)-(7) if $\Pi = 0$.

We will consider the marginal and conditional posterior densities under a flat prior and discuss existence conditions for these posteriors and their first and higher order moments. A summary of results is presented in Figure 1. For a description of the matrix normal and matrix t distributions we refer to Zellner (1971). The marginal posterior distributions of β and Π were derived by Drèze (1976, 1977) and Kleibergen and Van Dijk (1998), respectively. The contribution of this section is that it provides an overview of all the marginal and conditional posteriors, where it is discussed whether these are proper and whether these have finite moments. To the best of our knowledge, an analysis of the posterior moments of Π is novel. A warning is included that an unknowing user may use the Gibbs sampler in case of an improper posterior, while this would obviously not make sense. Further, for a concise derivation of these posteriors and their properties we refer to Appendix A.

The full conditional posterior distributions of β, Π and Σ are as follows:

- The **conditional posterior of Σ given β and Π** is easily seen from (5) as a kernel of the Inverse-Wishart density with T degrees of freedom and scale matrix $(u \;\; V)'(u \;\; V)$ with $u = y - X\beta$, $V = X - Z\Pi$.

- The **conditional posterior of β given Π and Σ** is the multivariate normal distribution $N(\mu_{\beta|\Pi,\Sigma}, \Omega_{\beta|\Pi,\Sigma})$, where $\mu_{\beta|\Pi,\Sigma} \equiv (X'X)^{-1}X'(y - \mu_{u|V,\Sigma})$ and $\Omega_{\beta|\Pi,\Sigma} \equiv \omega_{u|V,\Sigma}\,(X'X)^{-1}$; here we have $\mu_{u|V,\Sigma} \equiv V\Sigma_{22}^{-1}\sigma_{12}'$ and $\omega_{u|V,\Sigma} \equiv \sigma_{11} - \sigma_{12}\Sigma_{22}^{-1}\sigma_{12}'$.

- The **conditional posterior of Π given β and Σ** is the matrix normal distribution $N_{\mathrm{matrix}}(\mu_{\Pi|\beta,\Sigma}, \Omega_{V|u,\Sigma}, (Z'Z)^{-1})$ with $\mu_{\Pi|\beta,\Sigma} \equiv (Z'Z)^{-1}Z'(X - \mu_{V|u,\Sigma})$; here we have $\mu_{V|u,\Sigma} \equiv u\sigma_{11}^{-1}\sigma_{12}$ and $\Omega_{V|u,\Sigma} \equiv \Sigma_{22} - \sigma_{12}'\sigma_{11}^{-1}\sigma_{12}$. That is, $\mathrm{vec}(\Pi)|\beta, \Sigma, y, X, Z \sim N(\mathrm{vec}(\mu_{\Pi|\beta,\Sigma}), \Omega_{V|u,\Sigma} \otimes (Z'Z)^{-1})$.

Figure 1. Posterior distributions in the IV model with m possibly endogenous regressors, k instruments, and Gaussian errors, under a flat prior

<table>
<tr><td>Joint posterior</td><td>$p(\beta, \Pi, \Sigma \mid \text{data})$

Posterior has a ridge at $\Pi = 0$,
posterior is improper for $k \leq m$ and proper for $k > m$.</td></tr>
<tr><td>Conditional posteriors of β, Π and Σ</td><td>$p(\beta, \Pi, \Sigma \mid \text{data})$

↓ complete sum of squares in β → $\beta|\Pi, \Sigma, \text{data} \sim$ multivariate normal density

↓ complete sum of squares in Π → $\Pi|\beta, \Sigma, \text{data} \sim$ matrix normal density

↓ use properties of Inverse-Wishart distribution → $\Sigma|\beta, \Pi, \text{data} \sim \text{IW}(\Xi, T)$ where $\Xi = (u\ V)'(u\ V)$ for $u = y - X\beta$, $V = X - Z\Pi$

$p(\beta|\Pi, \Sigma, data)$, $p(\Pi|\beta, \Sigma, data)$ and $p(\Sigma|\beta, \Pi, data)$ are proper densities with finite moments for all values of β, Π and Σ in their domain, for any number of instruments $k \geq 1$ and for any number of possibly endogenous regressors $m \geq 1$.</td></tr>
<tr><td>Conditional posteriors of β and Π</td><td>$p(\beta, \Pi, \Sigma \mid \text{data})$
↓
Inverse-Wishart step on Σ in $p(\beta, \Pi, \Sigma \mid \text{data})$
↓
$p(\beta, \Pi \mid \text{data}) \propto \left|(u\ V)'(u\ V)\right|^{-T/2}$ for $u = y - X\beta$, $V = X - Z\Pi$

↓ $p(\Pi|\beta, \text{data}) \propto$ matrix t-density
If $p(\Pi|\beta, data)$ is proper for all values of β in its domain. The first few moments (i.e., at least up to the fourth moment) exist (if T is not very small).

↓ $p(\beta|\Pi, \text{data}) \propto$ multivariate t-density
If rank$(\Pi) < m$, then $p(\beta|\Pi, data)$ is improper. If rank$(\Pi) = m$, then $p(\beta|\Pi, data)$ is proper, and the first few moments (i.e., at least up to the fourth moment) exist (if T is not very small).</td></tr>
<tr><td>Marginal posteriors of β and Π</td><td>↓ t-density step on Π
↓
$p(\beta \mid \text{data}) \propto (u'u)^{-\frac{k}{2}} \left(\frac{u'M_Z u}{u'u}\right)^{\frac{T-k-m}{2}}$

For $p(\beta \mid data)$ is a t-density form times a polynomial (with $M_Z \equiv I - Z(Z'Z)^{-1}Z'$). Integer moments exist for order $r = 0, 1, 2, \ldots, k - m - 1$.

↓ t-density step on β
↓
$p(\Pi \mid y, X, Z) \propto |V'V|^{-\frac{T-1}{2}} |\Pi' Z' M_X Z \Pi|^{-\frac{1}{2}} \times \left(\frac{|\Pi' Z' M_X Z \Pi|}{|\Pi' Z' M_{(y\ X)} Z \Pi|}\right)^{\frac{T-m}{2}}$

For $p(\Pi \mid data)$ is a t-density form times by a rational function. If $k > m$, then the first few moments (i.e., at least up to the fourth moment) exist (if T is not very small).

Marginal posteriors of β and Π are improper for $k \leq m$ (exact or under-identification), and proper for $k > m$ (over-identification).</td></tr>
</table>

The conditional distributions of β and Π (after integrating out Σ) are as follows:

- The **conditional posterior of** β **given** Π is the multivariate t density with location vector $\hat{\beta}$ and scale matrix $s^2_{\hat{\beta}}(X'M_VX)^{-1}$ and $T-m$ degrees of freedom — where $M_\alpha \equiv I - \alpha\left(\alpha'\alpha\right)^{-1}\alpha'$, $\hat{\beta} \equiv \left(X'M_VX\right)^{-1}X'M_Vy$, and $s^2_{\hat{\beta}} \equiv (y-X\hat{\beta})'M_v(y-X\hat{\beta})/(T-m)$ — *given that* $(X'M_VX)$ has full rank m. The latter holds if and only if Π has full rank m. If $\text{rank}(\Pi) < m$, for example if $k < m$ (under-identification), then the conditional posterior of β given Π is improper.

- The **conditional posterior of** Π **given** β is a matrix t density with location matrix $\hat{\Pi}$, scale matrices $(Z'M_uZ)^{-1}$ and $S_{\hat{\Pi}}$, and $T-k-m+1$ degrees of freedom — with $\hat{\Pi} \equiv \left(Z'M_uZ\right)^{-1}Z'M_uX$ and $S_{\hat{\Pi}} = (X-Z\hat{\Pi})'M_u(X-Z\hat{\Pi}))$ — for any number of endogenous variables m, any number of instruments k and for every value of β.

The full conditional posteriors of β, Π and Σ are proper distributions for all values of β, Π and Σ in their domain, for any number of instruments $k \geq 1$ and for any number of possibly endogenous regressors $m \geq 1$. This implies that an unknowing user may *erroneously* apply the Gibbs sampler in case of exact identification (or even under-identification), even though the (joint) posterior distribution is improper, which will be discussed below. A Gibbs sampler that simulates only β and Π (after integrating out Σ) may also be *erroneously* applied in case of exact identification.

The marginal posterior distributions of β and Π are as follows:

- The **marginal posterior of** β is

$$p(\beta \mid y, X, Z) \propto (u'u)^{-\frac{T-m}{2}}\left(u'M_Zu\right)^{\frac{T-k-m}{2}}, \tag{8}$$

  which is a t-density multiplied by a polynomial, or

$$p(\beta \mid y, X, Z) \propto \left(\frac{u'M_Zu}{u'u}\right)^{\frac{T-k-m}{2}}(u'u)^{-\frac{k}{2}}, \tag{9}$$

  which is an improper density for $k \leq m$ (exact or under-identification), and a proper density for $k > m$ (over-identification). In the latter case, moments exist for (integer) order $r = 0, 1, 2, \ldots k-m-1$. From the marginal posterior of β, the conditional posterior of Π given β (which is proper for any β), and the conditional posterior of Σ given β and Π (which is proper for any β and Π) it is immediately clear that the joint posterior of β, Π and Σ is proper if and only if $k > m$ (over-identification).

- The **marginal posterior of** Π is

$$p\left(\Pi \mid y, X, Z\right) \propto |V'V|^{-\frac{T-1}{2}}\left|\Pi'Z'M_XZ\Pi\right|^{\frac{T-m-1}{2}}\left|\Pi'Z'M_{(y\ X)}Z\Pi\right|^{-\frac{T-m}{2}}, \tag{10}$$

a matrix t-density multiplied by a rational function, or

$$p\left(\Pi \mid y, X, Z\right) \propto |V'V|^{-\frac{T-1}{2}} \left(\frac{|\Pi' Z' M_X Z \Pi|}{\left|\Pi' Z' M_{(y\ X)} Z \Pi\right|} \right)^{\frac{T-m}{2}} |\Pi' Z' M_X Z \Pi|^{-\frac{1}{2}}, \qquad (11)$$

which is integrable only for $k > m$ (over-identification). In this case, the first few moments – i.e., at least up to the fourth moment – exist (given that T is not very small). For example, consider the case of $m = 1$. For $k = 1$ the factor $|\Pi' Z' M_X Z \Pi|^{-\frac{1}{2}}$ is not integrable around $\Pi = 0$, since $\int_{-1}^{1} \frac{1}{|\Pi|} d\Pi = \infty$. For $k = 2$ $|\Pi' Z' M_X Z \Pi|^{-\frac{1}{2}}$ is integrable around $\Pi = 0$, since $\int_{\{\Pi | \Pi_1^2 + \Pi_2^2 \leq 1\}} \frac{1}{(\Pi_1^2 + \Pi_2^2)^{1/2}} d\Pi = 2\pi$. Intuitively speaking, given that the posterior of Π is proper, higher order moments are finite, since problems regarding integrability only occur due to the 'vertical asymptote' for Π tending to values with $\text{rank}(\Pi) < m$, not due to fat tails (as for the posterior of β). Multiplying (11) by Π_{ij}^d $(i = 1, \ldots, k;\ j = 1, \ldots, m;\ d = 1, 2, \ldots)$ makes the function only 'easier' to integrate.

2.1 IV model with m possibly endogenous regressors under informative prior on β

If one specifies a proper prior $p(\beta)$ for β, e.g. a normal prior, so that

$$p(\beta, \Pi, \Sigma) \propto p(\beta) \times |\Sigma|^{-\frac{m+1}{2}}, \qquad (12)$$

then the marginal, conditional and joint posteriors in the IV model are always proper, no matter the dimensions k and m. The marginal posterior of β is then obviously obtained by multiplying (9) by $p(\beta)$:

$$p(\beta \mid y, X, Z) \propto p(\beta) \left(\frac{u' M_Z u}{u' u} \right)^{\frac{T-k-m}{2}} (u'u)^{-\frac{k}{2}}, \qquad (13)$$

whereas the conditional posteriors of Π given β, and of Σ given β and Π remain the same matrix t and Inverse-Wishart distributions. Finite prior moments of β then imply finite posterior moments of β (where the order of finite posterior moments may be $k - m + 1$ larger than the order of finite prior moments).

3 The potential of Direct Monte Carlo in IV models

Naturally, we should consider cases in which the posterior distribution is proper: therefore we consider the IV model (1)-(2) with $k \geq m + 1$ instruments (over-identification) under a flat prior, and also address the IV model with $k \geq m$ instruments (exact or over-identification) under a proper prior $p(\beta)$. For $m = 1$ or $m = 2$ the posterior moments of β can be computed accurately using quadrature. However, to analyze whether the instruments have explanatory power for the regressors or

whether the regressors are endogenous in the first place, one is often also interested in the posteriors of Π and Σ, respectively.

We propose the following *Acceptance-rejection within Direct Monte Carlo* (ARDMC) method, a simulation-consistent method for posterior simulation from the IV model with $m \geq 1$ possibly endogenous regressors:

Step 1: Draw β from its marginal posterior in (9) (or (13) under a proper prior on β), using the acceptance-rejection method (i.e., rejection sampling).

Step 2: Draw Π conditionally on β from its conditional matrix t posterior.

Step 3: Draw Σ conditionally on (β, Π) from its conditional Inverse-Wishart distribution.

The acceptance-rejection method in step 1 produces a set of independent drawings from the marginal posterior of β, which implies that we obtain a set of independent drawings of (β, Π, Σ). Obviously we only simulate Π and Σ for accepted drawings of β from step 1, so that steps 2 and 3 are exactly those of a DMC method, *directly* simulating independent drawings from the conditional distributions without a rejection step, an importance weighting step or Markov Chain steps. The independence of the drawings generated by ARDMC greatly helps the speed of convergence of simulation results, and facilitates the computation of accurate numerical standard errors or predictive likelihoods. Further, the computations can be easily performed in a parallelized fashion, which may yield another huge reduction of computing time on multiple core machines or computer clusters. We note that ARDMC is an analogue to the well-known "Metropolis-Hastings within Gibbs" sampling in the sense that one 'more difficult' step is used within an 'easier' simulation method.

Generally, the acceptance-rejection method has one major drawback: it requires a candidate density that provides a reasonably accurate approximation of the target (posterior) density and that dominates the target density (in the sense that the ratio of candidate over target has a finite maximum, that should be as small as possible). In this situation we are able to obtain such an accurate approximation for two reasons. First, the dimension m of β is typically low; e.g. $m = 1$ or $m = 2$ (although our ARDMC also appeared to work well in cases of $m = 4$). The lower dimension of β is also the reason why we approximate the marginal posterior of β rather than Π in step 1. Second, we use a novel adaptation of the *Mixture of t by Importance Sampling weighted Expectation Maximization* (MitISEM) method of Hoogerheide et al. (2012b). Until now the MitISEM procedure has only been used to construct an importance or candidate density for Importance Sampling or the independence chain Metropolis-Hastings algorithm. [2] Our novel adaptation aims at a high acceptance rate in the acceptance-rejection method rather than a low variance of the Importance Sampling weights. We use a mixture of Student-t densities as the

[2] The intimate link between Importance Sampling and the independence chain Metropolis-Hastings (MH) algorithm is pointed out by Liu (1996).

candidate, because it is easily evaluated, easily simulated from, and very flexible in the sense that it can approximate a wide variety of posterior shapes (e.g. multimodality or other types of non-elliptically curved shapes). Given a proper posterior kernel, the adapted MitISEM algorithm automatically finds an approximation of the posterior distribution: it starts with a Student-t distribution around the posterior mode and adds Student-t distributions as long as adding more Student-t components substantially increases the acceptance rate of the acceptance-rejection method. An IS weighted EM algorithm is used to optimize the locations, scales and degrees of freedom of all Student-t distributions, minimizing the Kullback-Leibler divergence between candidate and posterior. The allowed range of the degrees of freedom parameters of the Student-t distributions is restricted (from above) to ensure that the tails of the candidate distribution are fatter than those of the posterior, so that the candidate surely dominates the posterior.

In order to find the maximum of the ratio of the target density kernel to the candidate density, required for the acceptance-rejection sampling method, we proceed as follows. First, we compute this ratio for each of a large set of candidate draws for β (e.g., 100,000 candidate draws) that we will use in the acceptance-rejection sampling method, and find the value $\beta_{\text{arg max in sample}}$ that corresponds to the highest ratio. Second, we apply the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm with initial value $\beta_{\text{arg max in sample}}$ to find $\beta_{\text{arg max BFGS}}$ where the ratio takes a (local) maximum. In our examples below, β is 1-dimensional or 2-dimensional, so that it is relatively easy to check, using a graphical analysis and the evaluation of the ratio on a very fine grid of values for β, that the found $\beta_{\text{arg max BFGS}}$ is indeed the value $\beta_{\text{arg max global}}$ where the ratio takes its global maximum.

For higher dimensional β such a check would be more difficult. However, the above-mentioned two-step procedure followed by the application of the acceptance-rejection procedure is simulation-consistent, because the probability that the ratio takes its global maximum at the found value $\beta_{\text{arg max BFGS}}$ tends to 1 as the number of candidate draws tends to infinity. In other words, the procedure is the (simulation-consistent) regular acceptance-rejection method (with probability one) if the number of candidate draws tends to infinity. The reason for this is that the ratio is a smooth function of β, for which there exists a convex set B of values of β around $\beta_{\text{arg max global}}$ for which (i) the ratio is larger than in any point β outside B, (ii) the ratio is a concave function on B, (iii) the posterior probability $P_{\beta\in B}$ that β lies in B is positive, so that also the probability $\tilde{P}_{\beta\in B}$ that a candidate draw for β lies in B is positive. The latter implication holds true, since we know that the ratio has a finite global maximum $ratio_{max}$ (so that $\tilde{P}_{\beta\in B} \geq P_{\beta\in B}/ratio_{max} > 0$), which is ensured by the property that the candidate density has fatter tails than the posterior density.

That is, for large enough number of candidate draws $\beta_{\text{arg max in sample}}$ lies in B after which the BFGS method will yield $\beta_{\text{arg max BFGS}} = \beta_{\text{arg max global}}$. Note that if the ratio would have multiple global maxima, then the procedure will find one of these global maxima, which is sufficient for the acceptance-rejection method to work appropriately.

Summarizing, the proposed ARDMC method has the major advantages of DMC: it is fast and generates a set of independent drawings from the posterior. We will illustrate the quality of ARDMC for both simulated and empirical data sets, comparing its performance with the Gibbs sampler.

4 Applications of Acceptance-Rejection within DMC (ARDMC)

4.1 ARDMC simulation from posteriors for simulated data sets

We consider the posterior distribution in the IV model (1)-(2) under the flat prior for four simulated data sets. We consider the following cases of weak or strong instruments for either $m = 1$ or $m = 2$ possibly endogenous regressors:

Case 1: $k = 4$ weak instruments for $m = 1$ strongly endogenous regressor: $\Pi = 0.05\ \iota_{4\times 1}$, $\Sigma = \left(\begin{smallmatrix} 1 & 0.99 \\ 0.99 & 1 \end{smallmatrix}\right)$, $T = 50$, where ι denotes a vector or matrix of ones.

Case 2: $k = 6$ weak instruments for $m = 2$ strongly endogenous regressors: $\Pi = 0.1 \left(\begin{smallmatrix} \iota_{3\times 1} & 0_{3\times 1} \\ 0_{3\times 1} & \iota_{3\times 1} \end{smallmatrix}\right)$, $\Sigma = \left(\begin{smallmatrix} 1 & 0.99 & 0.99 \\ 0.99 & 1 & 0.99 \\ 0.99 & 0.99 & 1 \end{smallmatrix}\right)$, $T = 50$.

Case 3: $k = 4$ very strong instruments for $m = 1$ moderately endogenous regressor: $\Pi = \iota_{4\times 1}$, $\Sigma = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, $T = 100$.

Case 4: $k = 6$ very strong instruments for $m = 2$ moderately endogenous regressors: $\Pi = \left(\begin{smallmatrix} \iota_{3\times 1} & 0_{3\times 1} \\ 0_{3\times 1} & \iota_{3\times 1} \end{smallmatrix}\right)$, $\Sigma = \left(\begin{smallmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{smallmatrix}\right)$, $T = 100$.

For each case we take the instruments $z_t \sim N(0, I_k)$ i.i.d., and $\beta = 0_{m\times 1}$; the true value of $\beta = 0$ does not affect the shape of its posterior, only its location. Hoogerheide et al. (2007) considered posteriors (for cases of $m = 1$ endogenous regressor) on bounded regions: cases 1 and 3 are similar to the most extreme cases of Hoogerheide et al. (2007) where β has a highly non-elliptical (bimodal) posterior and an almost elliptical posterior, respectively.

Simulation results (without making use of parallelized computation or Rao-Blackwellization) for ARDMC and Gibbs sampling are reported in Table 1, where, to save space, for cases 2 and 4 (with $m = 2$) only simulation results are shown for β. In all cases ARDMC performs better than the Gibbs sampler: ARDMC requires substantially less computing time than Gibbs sampling (on an Intel Centrino$^{\text{tm}}$ processor) to yield similar precision (for very strong instruments) or much higher precision (for weak instruments). For ARDMC the Numerical Standard Error (NSE) of the estimated posterior mean is easily computed as the estimated posterior standard deviation divided by the square root of the number of accepted draws, whereas for

the Gibbs sampler we make use of the Initial Positive Sequence Estimator of Geyer (1992). The Effective Sample Size is the equivalent number of independent draws from the posterior that would lead to the same NSE; see Liu (2001).

For cases 1 and 2, the Gibbs sampler's ESS values are very low, less than 1000 for a set of 100000 draws. The slow movement of the Gibbs sequence through the parameter space is reflected by the high serial correlation for β and $\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$. Moreover, the difference in quality between ARDMC and Gibbs sampling is even larger than suggested by the NSE and ESS. The Gibbs sampler misses relevant parts of the parameter space. For case 1 (2), Figure 2 (4) shows that the posterior of β is bimodal; that the adapted MitISEM method generates (in merely 24 (11) seconds) a mixture of 7 (3) Student-t distributions that provides a reasonably accurate approximation of the posterior, which yields a set of 100000 candidate draws of which 64133 (47157) independent posterior draws are accepted; and that the Gibbs sequence is stuck in a region of parameter values around one of the two modes. Even if we generate ten million draws, the Gibbs sequence is still stuck in the region around one of the two modes. For case 1, the histogram of each 100th Gibbs draw in Figure 3 shows that one of the two modes is now completely 'covered', but that the other mode is still 'missed' by ten million consecutive Gibbs draws. So, reliable Gibbs sampling requires a chain that is impracticably long in this example, because the Gibbs sampler is very poorly mixing. Note that we do not claim that the Gibbs sampler is (theoretically) nonergodic; for an infinite number of draws, the Gibbs sampler will move between the modes. For case 2 the bottom-right panel of Figure 4 shows that all Gibbs draws are from the bottom-left "mountain" around the origin, which is much smaller than the top-right "mountain" around $(\beta_1, \beta_2) = (1, 1)$ (in the sense that the first contains much less posterior probability mass). Therefore this bottom-left "mountain" may look a bit 'wider' for the Gibbs sampler than for ARDMC in the scatter plots of Figure 4, since only a small fraction of the ARDMC draws is located in this area.

To explain the posterior shapes for case 2, we rewrite the marginal posterior of β as:

$$\begin{aligned} p(\beta \mid y, X, Z) &\propto \left(\frac{(y - X\beta)' M_Z (y - X\beta)}{(y - X\beta)'(y - X\beta)} \right)^{\frac{T-k-m}{2}} ((y - X\beta)'(y - X\beta))^{-\frac{k}{2}} \\ &= \left(1 - \frac{(P_Z y)' M_{P_Z X} (P_Z y) + (\beta - \hat{\beta}_{2SLS})' X' P_Z X (\beta - \hat{\beta}_{2SLS})}{y' M_X y \qquad + (\beta - \hat{\beta}_{OLS})' X' X (\beta - \hat{\beta}_{OLS})} \right)^{\frac{T-k-m}{2}} \times \\ &\quad (y' M_X y + (\beta - \hat{\beta}_{OLS})' X' X (\beta - \hat{\beta}_{OLS}))^{-\frac{k}{2}} \end{aligned} \tag{14}$$

with $P_Z = Z(Z'Z)^{-1}Z'$, $\hat{\beta}_{2SLS} = (X' P_Z X)^{-1} X' P_Z y$, $\hat{\beta}_{OLS} = (X'X)^{-1} X' y$.

In case of weak instruments $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$ may be relatively far apart. Indeed for our case 2 the difference between $\hat{\beta}_{OLS} = (0.44, 0.58)'$ and $\hat{\beta}_{2SLS} = (0.70, 0.80)'$ is relatively large, and $(P_Z y)' M_{P_Z X} (P_Z y)$ and $y' M_X y$ are both small, due to the weakness of the instruments and the strong endogeneity, respectively. This implies

that (14) is very low for $\beta \approx \hat{\beta}_{OLS}$, whereas on both sides of β_{OLS} there are regions where (14) is not negligible. $X'X$ has eigenvectors $(-0.72, 0.70)$ and $(0.70, 0.72)$ with eigenvalues 2.57 and 91.58, respectively, which explains the 'ravine' of low posterior density values around the line $\{\beta = \hat{\beta}_{OLS} + \lambda(-0.72, 0.70)' | \lambda \in \mathbb{R}\}$. The top row of Figure 5 shows contour plots of the logarithm of the posterior, the numerator $(P_Z y)' M_{P_Z X}(P_Z y) + (\beta - \hat{\beta}_{2SLS})' X' P_Z X (\beta - \hat{\beta}_{2SLS})$, and the denominator $y' M_X y + (\beta - \hat{\beta}_{OLS})' X' X (\beta - \hat{\beta}_{OLS})$ of the ratio in (14). For case 1 the (bimodal) posterior shapes are explained in an analogous fashion.

For cases 3 and 4 the Gibbs sampler performs reasonably well, although it requires more computing time than ARDMC. It should be noted that the simulated instruments are very strong; in many empirical applications – of which one will be considered below – the elements of Π are less significant and more similar among columns, or the instruments are correlated. In cases 3 and 4 the posterior is closer to an elliptical distribution, see Figure 6. Therefore it takes the adapted MitISEM method even less time to construct an approximation of the posterior, since mixtures of only two Student-t distributions are used. Figure 6 shows that, although its (far) tails are Student-t type, the 'middle part' of the posterior of β in case 3 is more like a Gaussian distribution. The bottom row of Figure 5 shows the logarithm of the posterior of β for case 4. Here the difference between $\hat{\beta}_{OLS} = (0.04, 0.19)'$ and $\hat{\beta}_{2SLS} = (-0.02, 0.14)'$ is smaller than for case 2. Moreover, the strong instruments and moderate endogeneity imply that $(P_Z y)' M_{P_Z X}(P_Z y)$ and $y' M_X y$ are much larger than in case 2, so that there is no 'ravine' through $\hat{\beta}_{OLS}$. Far away from the posterior mode, the shapes (driven by the eigenvectors and eigenvalues of $X' P_Z X$ and $X'X$) become somewhat similar to case 4, but this occurs only for very low levels of the posterior density. For case 3 the posterior shapes are explained in an analogous fashion.

Table 2 shows simulation results for case 1 and 3 where we make use of Rao-Blackwellization in order to estimate the posterior means (and standard deviations) of β and Π. The benefits of Rao-Blackwellization depend crucially on three factors: (i) the simulation method, (ii) the strength of the instruments, and (iii) the parameter that is considered. For ARDMC the benefits are substantial unless one considers the posterior mean of β in case 1 of weak instruments. For the Gibbs sampler the benefits are substantial unless one considers the posterior mean of either β or Π in case 1 of weak instruments. In the latter case, one faces negligible gains.

If we estimate the posterior mean of β, then the benefits of Rao-Blackwellization stem from the fact that the standard deviation of $E[\beta | \Pi, \Sigma, data]$ (where a posterior draw is used for (Π, Σ)), reported in the last column of Table 2, is smaller than the posterior standard deviation of β itself, since Rao-Blackwellization means that we compute the average of the first instead of the second. However, in our case 1 of weak instruments the benefits are small, since the standard deviation of $E[\beta | \Pi, \Sigma, data]$ almost equals the posterior standard deviation of β due to the strong posterior dependence between β, Π, and Σ. For Rao-Blackwellization of the Gibbs sampler, a

disadvantage is that the serial correlation in the Gibbs sequence may be much larger for $E[\beta|\Pi, \Sigma, data]$ than for β itself, so that the benefits due to the smaller standard deviation are partly (or in certain cases almost completely) lost. This phenomenon is observed for the estimation of the posterior mean of Π in case 1 of weak instruments. For ARDMC this disadvantage is not present, since ARDMC generates a set of independent draws. Therefore, the relative benefits of Rao-Blackwellization are larger for ARDMC than for the Gibbs sampler. In fact, in our case 1 of weak instruments the difference in precision between ARDMC and Gibbs is much larger than the benefits from Rao-Blackwellization in either procedure.

For the Gibbs sampler in case 1 the reported ESS for Π_1 and Π_3 is even slightly smaller for the case with Rao-Blackwellization than without Rao-Blackwellization. This is merely caused by the fact that the used NSE and posterior standard deviation are estimates.

We have also considered cases with $m = 4$, with weak or strong instruments, similar to cases 2 and 4. Then we still observe a similar 'victory' of ARDMC (with acceptance rates of 38% and 52%) over the Gibbs sampler in terms of numerical accuracy. This is no surprise, as Hoogerheide et al. (2012b) show examples of posteriors with 17 and 36 parameters, where the MitISEM method provides importance densities that are reasonably accurate approximations of the joint posterior.

Table 1
Simulation results for posterior distributions in IV model for four simulated data sets (without making use of Rao-Blackwellization)

	ARDMC			Gibbs sampling				
	posterior			posterior				
	mean	st.dev.	NSE	mean	st.dev.	NSE	s.c.	ESS
Case 1 ($k = 4$ weak instruments for $m = 1$ strongly endogenous regressor):								
β	1.2984	0.6318	0.0025	1.4661	0.2265	0.0189	0.995	143
Π_1	-0.0493	0.0733	0.0003	-0.0753	0.0476	0.0016	0.245	908
Π_2	-0.1049	0.1111	0.0004	-0.1461	0.0677	0.0038	0.667	317
Π_3	-0.0886	0.1075	0.0004	-0.1303	0.0566	0.0030	0.573	368
Π_4	-0.1167	0.1218	0.0005	-0.1615	0.0749	0.0042	0.684	317
$\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$	-0.6482	0.7201	0.0028	-0.9659	0.0273	0.0015	0.815	354
number of draws	100000	candidate draws		100000 (+1000 burnin)				
	64133	accepted draws						
computing time: $*$ total	44 s			61 s				
$*$ candidate	24 s							
$*$ sampling	20 s							
Case 2 ($k = 6$ weak instruments for $m = 2$ strongly endogenous regressors):								
β_1	0.5814	0.3666	0.0017	0.1634	0.3285	0.0114	0.970	832
β_2	0.7077	0.3432	0.0016	0.3340	0.3384	0.0138	0.973	595
number of draws	100000	candidate draws		100000 (+1000 burnin)				
	47153	accepted draws						
computing time: $*$ total	47 s			82 s				
$*$ candidate	11 s							
$*$ sampling	36 s							
Case 3 ($k = 4$ very strong instruments for $m = 1$ moderately endogenous regressor):								
β	-0.0174	0.0514	$0.1886 \cdot 10^{-3}$	-0.0177	0.0512	$0.1985 \cdot 10^{-3}$	0.358	66580
Π_1	0.9893	0.0888	$0.3258 \cdot 10^{-3}$	0.9888	0.0888	$0.2901 \cdot 10^{-3}$	0.059	93807
Π_2	1.0412	0.0845	$0.3101 \cdot 10^{-3}$	1.0410	0.0843	$0.2849 \cdot 10^{-3}$	0.103	87654
Π_3	0.9921	0.0864	$0.3170 \cdot 10^{-3}$	0.9918	0.0863	$0.2798 \cdot 10^{-3}$	0.061	95172
Π_4	1.0057	0.0868	$0.3185 \cdot 10^{-3}$	1.0063	0.0864	$0.2885 \cdot 10^{-3}$	0.068	89414
$\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$	0.4498	0.0898	$0.3298 \cdot 10^{-3}$	0.4502	0.0897	$0.3193 \cdot 10^{-3}$	0.206	78861
number of draws	100000	candidate draws		100000 (+1000 burnin)				
	74224	accepted draws						
computing time: $*$ total	32 s			64 s				
$*$ candidate	7 s							
$*$ sampling	25 s							
Case 4 ($k = 6$ very strong instruments for $m = 2$ moderately endogenous regressors):								
β_1	-0.0305	0.0532	$0.2170 \cdot 10^{-3}$	-0.0303	0.0533	$0.2303 \cdot 10^{-3}$	0.424	53473
β_2	0.1290	0.0480	$0.1956 \cdot 10^{-3}$	0.1296	0.0479	$0.1966 \cdot 10^{-3}$	0.383	59241
number of draws	100000	candidate draws		100000 (+1000 burnin)				
	60105	accepted draws						
computing time: $*$ total	41 s			85 s				
$*$ candidate	7 s							
$*$ sampling	34 s							

NSE = Numerical Standard Error of estimated posterior mean

s.c. = first order serial correlation in Gibbs sequence

ESS = Effective Sample Size (for estimating the posterior mean)

Table 2
Simulation results for posterior distributions in IV model for simulated data sets using Rao-Blackwellization

	ARDMC					Gibbs sampling					
	posterior				st.dev.	posterior					st.dev.
	mean	st.dev.	NSE	ESS	cond. mean	mean	st.dev.	NSE	s.c.	ESS	cond. mean
Case 1 ($k = 4$ weak instruments for $m = 1$ strongly endogenous regressor):											
β	1.2984	0.6318	0.0025	64 173	0.6316	1.4662	0.2266	0.0189	0.995	144	0.2260
Π_1	-0.0495	0.0736	0.0002	90 713	0.0618	-0.0753	0.0477	0.0016	0.850	898	0.0240
Π_2	-0.1049	0.1111	0.0004	72 580	0.1042	-0.1462	0.0679	0.0038	0.982	324	0.0555
Π_3	-0.0885	0.1075	0.0004	72 074	0.1014	-0.1305	0.0567	0.0030	0.945	364	0.0429
Π_4	-0.1166	0.1216	0.0005	72 113	0.1147	-0.1618	0.0750	0.0042	0.979	317	0.0622
Case 3 ($k = 4$ very strong instruments for $m = 1$ moderately endogenous regressor):											
β	-0.0175	0.0512	$0.1127 \cdot 10^{-3}$	206 468	0.0307	-0.0176	0.0512	$0.1199 \cdot 10^{-3}$	0.381	182 287	0.0307
Π_1	0.9889	0.0888	$0.0821 \cdot 10^{-3}$	1 170 614	0.0244	0.9888	0.0888	$0.0875 \cdot 10^{-3}$	0.373	1 029 485	0.0224
Π_2	1.0408	0.0846	$0.0994 \cdot 10^{-3}$	723 466	0.0271	1.0407	0.0846	$0.1067 \cdot 10^{-3}$	0.397	627 583	0.0272
Π_3	0.9918	0.0861	$0.0777 \cdot 10^{-3}$	1 230 014	0.0212	0.9917	0.0861	$0.0749 \cdot 10^{-3}$	0.174	1 321 586	0.0212
Π_4	1.0064	0.0867	$0.0814 \cdot 10^{-3}$	1 135 037	0.0222	1.0063	0.0867	$0.0863 \cdot 10^{-3}$	0.359	1 009 336	0.0222

NSE = Numerical Standard Error of estimated posterior mean

s.c. = first order serial correlation in Gibbs sequence of conditional posterior mean

ESS = Effective Sample Size (for estimating the posterior mean)

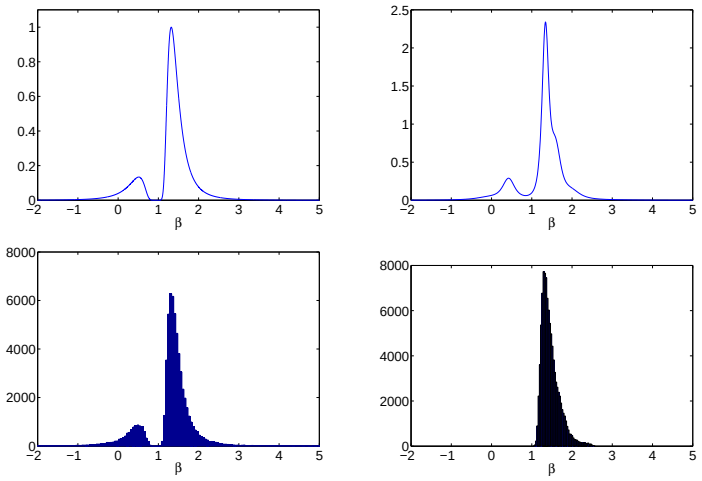

Figure 2. Marginal posterior of β in case 1 ($k = 4$ weak instruments for $m = 1$ strongly endogenous regressor): posterior density kernel (top left); candidate density (mixture of 7 Student-t densities) (top right); histogram of ARDMC draws (bottom left); histogram of Gibbs draws (bottom right).

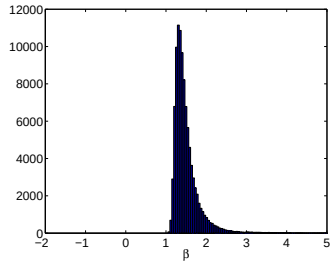

Figure 3. Marginal posterior of β in case 1 ($k = 4$ weak instruments for $m = 1$ strongly endogenous regressor): histogram of each 100th draw in a Gibbs sequence of ten million draws.

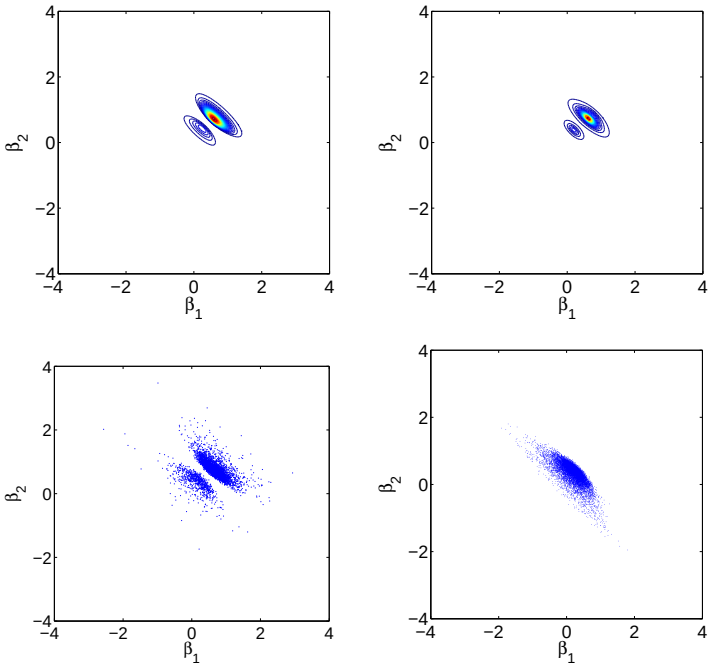

Figure 4. Marginal posterior of β in case 2 ($k = 6$ weak instruments for $m = 2$ strongly endogenous regressors): contour plot of posterior density kernel (top left); contour plot of candidate density (mixture of 3 Student-t densities) (top right); scatter plot of posterior draws generated by ARDMC (bottom left); scatter plot of posterior draws generated by the Gibbs sampler (bottom right).

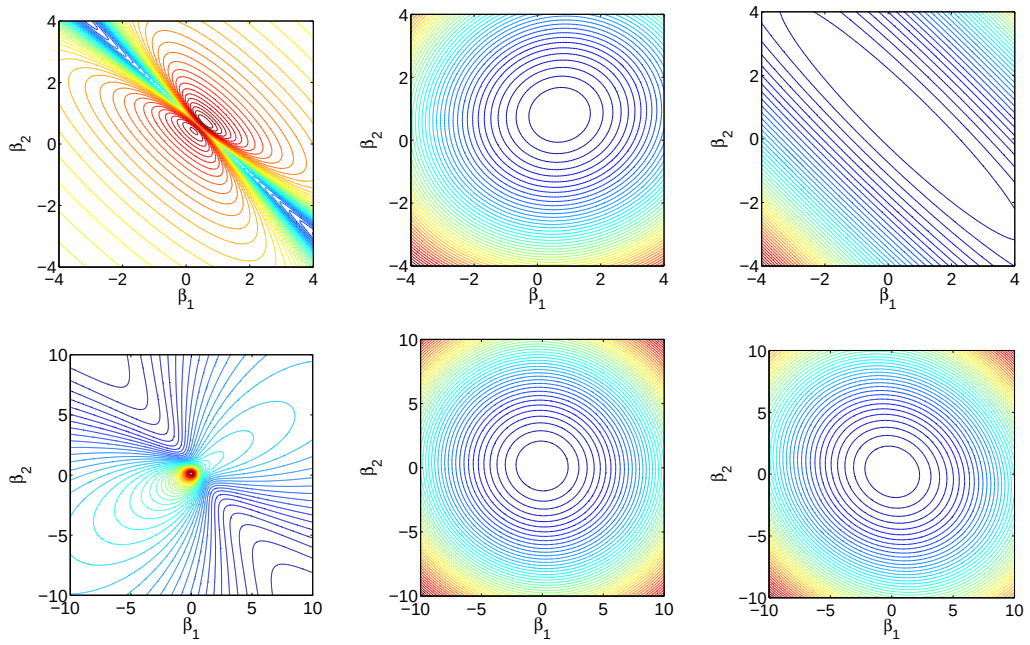

Figure 5. Contour plots: logarithm of posterior density kernel (left); numerator $(P_Z y)' M_{P_Z X}(P_Z y) + (\beta - \hat{\beta}_{2SLS})' X' P_Z X (\beta - \hat{\beta}_{2SLS})$ *of ratio in* (14) *(middle); denominator* $y' M_X y + (\beta - \hat{\beta}_{OLS})' X' X (\beta - \hat{\beta}_{OLS})$ *of ratio in* (14) *(right).*
Top row: case 2 ($k = 6$ *weak instruments for* $m = 2$ *strongly endogenous regressors).*
Bottom row: case 4 ($k = 6$ *very strong instruments for* $m = 2$ *moderately endogenous regressors).*

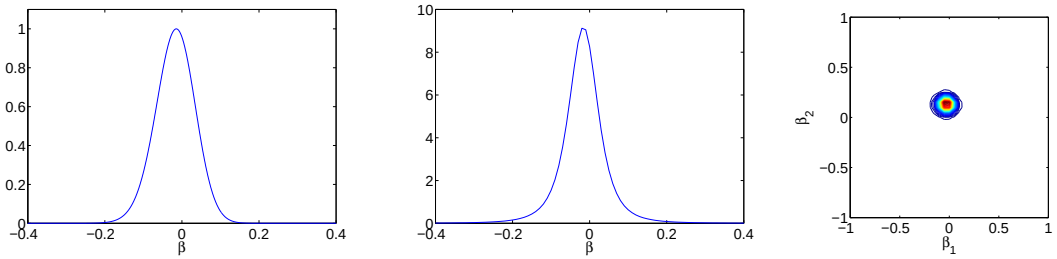

Figure 6. Marginal posterior of β *in case 3 (*$k = 4$ *strong instruments for* $m = 1$ *moderately endogenous regressor): posterior density kernel (left); candidate density (mixture of 2 Student-t densities) (middle).*
Marginal posterior of β *in case 4 (*$k = 6$ *strong instruments for* $m = 2$ *moderately endogenous regressors): contour plot of posterior density kernel (right).*

Our first empirical example is due to Acemoglu et al. (2001), see also Conley et al. (2008). Acemoglu et al. (2001) consider the effect of the risk of expropriation on the GDP per capita. To solve the endogeneity problem, European settler mortality is used as an instrument for the risk of expropriation. The idea behind this instrument is that in former colonies with high settler mortality Europeans could not settle and, therefore, set up more extractive institutions. The sample consists of $T = 64$ ex-colony countries. The model is given by

$$\log \mathrm{GDP}_t = \mathrm{APER}_t \, \beta + w_t \gamma_1 + u_t, \tag{15}$$

$$\mathrm{APER}_t = \log \mathrm{mortality}_t \, \Pi + w_t \gamma_2 + v_t, \tag{16}$$

where the dependent variable $\log \mathrm{GDP}_t$ is the logarithm of GDP per capita in 1995, the $m = 1$ possibly endogenous regressor APER_t is the 'Average protection against expropriation risk' for the period 1985-1995, the $k = 1$ instrument $\log \mathrm{mortality}_t$ is the logarithm of European settler mortality. The exogenous regressors w_t are the conditioning variables including a constant, latitude and dummies for African and Asian countries. The data y_t, x_t and z_t in (1)-(2), obtained as the residuals of $\log \mathrm{GDP}_t$, APER_t and $\log \mathrm{mortality}_t$ after regression on the control variables w_t, are shown in Figure 7. The left panel illustrates the weakness of the instrument.

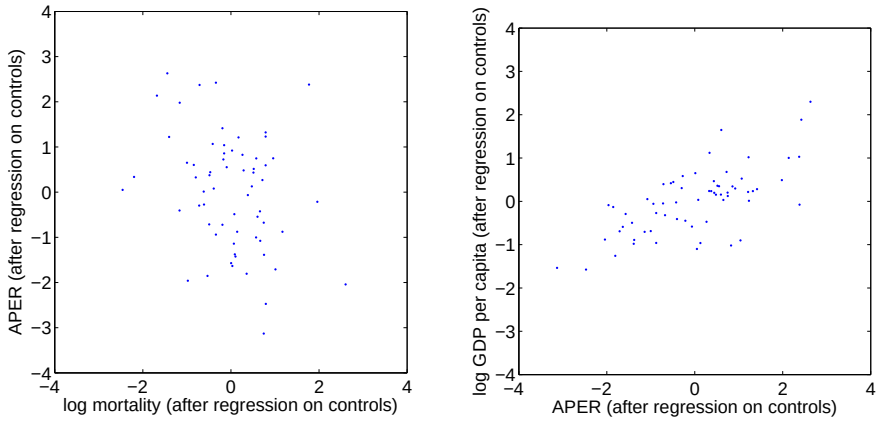

Figure 7. Empirical data set of Acemoglu et al. (2001) that is used in the first empirical example.

In this case of exact identification ($k = m = 1$) we use a proper, non-informative prior for β, a normal distribution with mean 0 and standard deviation 100. Under the flat prior, the posterior would be improper.

Simulation results are reported in Table 3. The posterior mean and standard deviation of Π show that this concerns a case of weak instruments, similar to case 1 of the simulated data sets. Figure 8 shows that the Gibbs sampler, which suffers from a huge serial correlation in the Gibbs sequence of drawings, misses a relevant part (consisting of negative values of β) of the bimodal posterior. On the other hand, ARDMC yields (with a high acceptance rate) a set of posterior drawings in a quick

and reliable fashion.

Our second empirical example uses a data set that is made available by the German Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW), Berlin. For more information about the SOEP, we refer to Wagner et al. (1993, 2007). The data set has been used by Hoogerheide et al. (2012a). The sample consists of a cross section of $T = 8244$ individuals (without missing values) in the year 2004. The model is given by

$$\log \text{wage}_t = \text{education}_t \, \beta + w_t \gamma_1 + u_t, \tag{17}$$

$$\text{education}_t = \text{father's education}_t \, \Pi + w_t \gamma_2 + v_t, \tag{18}$$

where the dependent variable $\log \text{wage}_t$ is the logarithm of hourly wage in 2004, the $m = 1$ possibly endogenous regressor education_t is the number of years of education, the $k = 3$ instruments ($\text{father's education}_t$) are dummy variables reflecting father's secondary education: 'Hauptschule' (9 years), 'Realschule' (10 years), 'Fachhochschulreife' (12 years) or 'Abitur' (13 years). We take 'Hauptschule' as the reference category. The exogenous regressors w_t are a constant, respondent's labor market experience (in its linear and squared terms), gender, wealth (as proxied by the respondent's income from assets), marriage status, nationality, whether the respondent lives in the former West-Germany, whether the respondent is self-employed, industry dummies, and the duration that an individual has been unemployed in his or her entire working life. The data y_t, x_t and z_t in (1)-(2) are obtained as the residuals of $\log \text{wage}_t$, education_t and $\text{father's education}_t$ after regression on the control variables w_t. In this case of over-identification ($k = m + 2$) a flat prior would imply a proper posterior with a finite posterior mean of β. However, since we are also interested in the posterior variance of β, and particularly since we desire to compare (finite) Numerical Standard Errors for the estimated posterior mean of β, we specify a proper, non-informative prior for β, a standard normal distribution.

Simulation results are reported in Table 3. The posterior means and standard deviations of the elements of Π show that this concerns a case of strong instruments, similar to case 3 of the simulated data sets. However, although the estimated posterior moments are similar for Gibbs sampling and ARDMC, the Gibbs sampler's Effective Sample Size for estimating the posterior mean of β and ρ is merely around 6500 (for 100000 draws), due to the rather high serial correlation in the Gibbs sequences of draws of β and ρ. Therefore, ARDMC clearly provides a higher numerical accuracy than the Gibbs sampler in this case of strong instruments.

We now estimate a model with $m = 2$ possibly endogenous regressors:

$$\log \text{wage}_t = \text{education}_t \, \beta_1 + \text{unemployment}_t \beta_2 + w_t \gamma_1 \; + u_t, \tag{19}$$

$$\text{education}_t = \text{father's education}_t \, \Pi_1 \qquad + w_t \gamma_{21} + v_{t1}, \tag{20}$$

$$\text{unemployment}_t = \text{father's education}_t \, \Pi_2 \qquad + w_t \gamma_{22} + v_{t2}, \tag{21}$$

where both years of education and the duration that an individual has been unem-

ployed (in his or her entire working life) are considered as possibly endogenous regressors, where the latter is now also excluded from w_t. The idea behind this choice is that unemployment duration, just like education, may be correlated with a latent 'ability' that affects the error term u_t in (19). We specify a proper, non-informative $N(0, I_2)$ prior for β. Simulation results are reported in Table 3. Here the Gibbs sequence of β has a huge serial correlation, illustrated by Figure 9, which causes a very low ESS. On the other hand, ARDMC still has a high acceptance rate of approximately 60%. This model provides a nice example of a marginal posterior of β that is rather close to an elliptical distribution, where the Gibbs sampler would require many more drawings (and much more computing time) than ARDMC to yield accurate estimates of the posterior moments.

Table 4 shows simulation results for the first two empirical models where we make use of Rao-Blackwellization in order to estimate the posterior means (and standard deviations) of β and Π. The findings are similar to those for the simulated data sets. In the first case of one weak instrument the difference in precision between ARDMC and Gibbs is much larger than the benefits from Rao-Blackwellization in either procedure. In the second case of three strong instruments, the benefits from Rao-Blackwellization are substantial, where the relative benefits are even larger for ARDMC than for the Gibbs sampler.

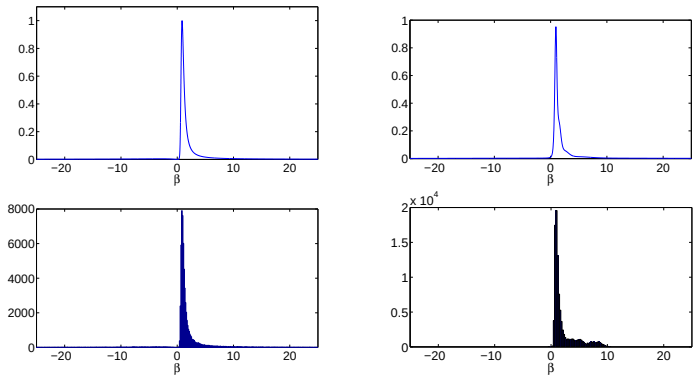

Figure 8. Marginal posterior of β for example due to Acemoglu et al. (2001): posterior density kernel (top left); candidate density (mixture of 8 Student-t densities) (top right); histogram of ARDMC draws (bottom left); histogram of Gibbs draws (bottom right).

Table 3
Posterior simulation results for empirical examples (without making use of Rao-Blackwellization)

	ARDMC			Gibbs sampling				
	posterior			posterior				
	mean	st.dev.	NSE	mean	st.dev.	NSE	s.c.	ESS
Example ($m = 1$, $k = 1$, $T = 64$ countries) due to Acemoglu et al. (2001) under proper, noninformative N(0,100) prior on β: y_t = log GDP per capita; x_t = Average Protection against Expropriation Risk; z_t = log European settler mortality.								
β	1.7936	26.6797	0.0991	2.2131	2.1621	0.5358	0.999	16
Π_1	-0.2564	0.2066	0.0008	-0.2899	0.1898	0.0302	0.804	39
$\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$	-0.6613	0.5868	0.0022	-0.8562	0.1593	0.0178	0.932	81
number of draws	100000	candidate draws		100000 (+1000 burnin)				
	72472	accepted draws						
computing time: $*$ total	48 s			59 s				
$*$ candidate	27 s							
$*$ sampling	21 s							
Example ($m = 1$, $k = 3$, $T = 8244$ individuals) of German SOEP data under proper, noninformative $N(0, 1)$ prior on β: y_t = log hourly wage; x_t = years of education; z_t = dummy variables indicating father's education.								
β	0.0812	0.0064	$0.2196 \cdot 10^{-4}$	0.0813	0.0063	$0.7903 \cdot 10^{-4}$	0.892	6435
Π_1	1.1525	0.0738	$2.5447 \cdot 10^{-4}$	1.1524	0.0734	$2.3373 \cdot 10^{-4}$	0.001	98490
Π_2	1.4388	0.3095	$10.6693 \cdot 10^{-4}$	1.4387	0.3112	$10.1407 \cdot 10^{-4}$	0.002	94169
Π_3	2.4151	0.0801	$2.7604 \cdot 10^{-4}$	2.4156	0.0800	$2.7242 \cdot 10^{-4}$	0.013	86191
$\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$	-0.0858	0.0331	$1.1421 \cdot 10^{-4}$	-0.0859	0.0331	$4.1193 \cdot 10^{-4}$	0.891	6458
number of draws	100000	candidate draws		100000 (+1000 burnin)				
	84156	accepted draws						
computing time: $*$ total	38 s			68 s				
$*$ candidate	10 s							
$*$ sampling	28 s							
Example ($m = 2$, $k = 3$, $T = 8244$ individuals) of German SOEP data under proper, noninformative $N(0, I_2)$ prior on β: y_t = log hourly wage; x_t = years of education, unemployment duration; z_t = dummy variables indicating father's education.								
β_1	0.0653	0.0196	$0.0798 \cdot 10^{-3}$	0.0623	0.0188	$4.3665 \cdot 10^{-3}$	0.988	19
β_2	-0.2595	0.2380	$0.9692 \cdot 10^{-3}$	-0.3030	0.2298	$57.8415 \cdot 10^{-3}$	0.999	16
number of draws	100000	candidate draws		100000 (+1000 burnin)				
	60312	accepted draws						
computing time: $*$ total	56 s			91 s				
$*$ candidate	8 s							
$*$ sampling	48 s							

NSE, s.c., ESS: see Table 1.

Table 4
Posterior simulation results for empirical examples using Rao-Blackwellization

	ARDMC					Gibbs sampling					
	posterior				st.dev.	posterior					st.dev.
	mean	st.dev.	NSE	ESS	cond. mean	mean	st.dev.	NSE	s.c.	ESS	cond. mean
Example ($m = 1$, $k = 1$, $T = 64$ countries) due to Acemoglu et al. (2001) under proper, noninformative N(0,100) prior on β: y_t = log GDP per capita; x_t = Average Protection against Expropriation Risk; z_t = log European settler mortality.											
β	1.7937	26.6794	0.0991	72 471	26.6795	2.2130	2.1619	0.5357	0.999	16	2.1613
Π_1	-0.2561	0.2068	0.0008	85 266	0.1906	-0.2893	0.1896	0.0302	0.986	39	0.1699
Example ($m = 1$, $k = 3$, $T = 8244$ individuals) of German SOEP data under proper, noninformative $N(0, 1)$ prior on β: y_t = log hourly wage; x_t = years of education; z_t = dummy variables indicating father's education.											
β	0.0813	0.0064	$0.2073 \cdot 10^{-4}$	94 312	0.0060	0.0813	0.0064	$0.7475 \cdot 10^{-4}$	0.892	7 237	0.0060
Π_1	1.1522	0.0734	$0.0708 \cdot 10^{-4}$	107 500 475	0.0021	1.1522	0.0734	$0.1797 \cdot 10^{-4}$	0.683	16 690 268	0.0020
Π_2	1.4382	0.3113	$0.4583 \cdot 10^{-4}$	46 134 501	0.0133	1.4382	0.3113	$1.6559 \cdot 10^{-4}$	0.909	3 534 037	0.0132
Π_3	2.4153	0.0799	$0.2731 \cdot 10^{-4}$	8 556 410	0.0079	2.4153	0.0799	$0.9337 \cdot 10^{-4}$	0.887	731 870	0.0078

NSE, s.c., ESS: see Table 2.

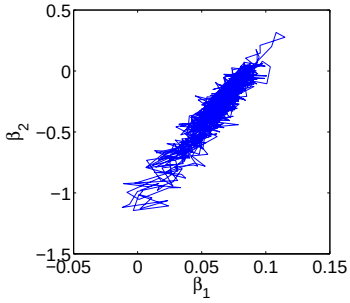

Figure 9. IV model with $m = 2$ possibly endogenous regressors (education and unemployment spell): scatter plot of every 100th draw in the Gibbs sequence

5 Further possibilities of ARDMC

Suppose that one is also interested in the effect of the included exogenous variables w_t $(1 \times n)$ in

$$y_t = x_t\beta + w_t\gamma_1 + u_t, \tag{22}$$
$$x_t = z_t\Pi + w_t\Gamma_2 + v_t, \tag{23}$$

where one specifies the prior $p(\beta, \Pi, \Sigma, \gamma_1, \Gamma_2) \propto p(\beta)\,|\Sigma|^{-\frac{m+2}{2}}$. After applying the ARDMC method to the posteriors where y_t, x_t and z_t are replaced by their residuals after regression on w_t (and where T is replaced by $T-n$), the ARDMC procedure is easily extended by a fourth step where $(\gamma_1, \text{vec}(\Gamma_2))'$ is simulated from its conditional posterior (conditional upon β, Π and Σ), $N(\text{vec}[(W'W)^{-1}W'(y - X\beta)\ \ (W'W)^{-1}W'(X - Z\Pi)], \Sigma \otimes (W'W)^{-1})$ with $W = (w_1', \ldots, w_T')'$.

Further, the ARDMC procedure can be applied in a non-linear IV model

$$y_t = f(x_t, \beta) + u_t, \tag{24}$$
$$x_t = z_t\Pi + v_t, \tag{25}$$

with $(u_t, v_t')' \sim NID(0_{(m+1)\times 1}, \Sigma)$ under the prior $p(\beta, \Pi, \Sigma) \propto p(\beta)\,|\Sigma|^{-\frac{m+2}{2}}$. Then the adapted MitISEM method in step 1 will aim at the marginal posterior of β

$$p(\beta \mid y, X, Z) \propto p(\beta)\frac{((y - f(X,\beta))'M_Z(y - f(X,\beta)))^{\frac{T-k-m}{2}}}{((y - f(X,\beta))'(y - f(X,\beta)))^{\frac{T-m}{2}}} \tag{26}$$

where $f(X, \beta) \equiv (f(x_1, \beta), \ldots, f(x_T, \beta))'$. For example, a possibly non-linear effect of education on the logarithm of income could be investigated by specifying $f(x_t, \beta) = \beta_0 + \beta_1 x_t^{\beta_2}$. The conditional posteriors (of Π given β, and of Σ given β and Π) in steps 2 and 3 simply remain the matrix t and inverse-Wishart distributions (with $u \equiv y - f(X, \beta)$).

The ARDMC method can not be readily applied to the posterior under Jeffreys' prior. Although Jeffreys' prior eliminates the vertical asymptote of the marginal posterior of Π around $\Pi = 0$, the marginal posterior of β (and the posterior of (β, Π)) may still be highly non-elliptical; see subsection 4.2.2 of Hoogerheide (2006). Moreover, posterior moments of β and Π do not exist, since the posteriors of β and Π under Jeffreys' prior have Cauchy-type tails, even in case of over-identification. Summarizing, not all the issues due to local non-identification are solved by the use of Jeffreys' prior, and it leads to posterior properties that may be found undesirable. For posterior simulation under Jeffreys' prior we refer to the methods developed by Kleibergen and Van Dijk (1998) and Kleibergen and Paap (2002).

6 Conclusions and Future Work

We discussed Bayesian inferential procedures within the instrumental variables regression model and focused on two issues: existence conditions for posterior moments under a flat prior and the potential of Direct Monte Carlo (DMC) approaches for efficient evaluation of such possibly highly non-elliptical posteriors. We discussed that, for the general case of m endogenous variables, posterior moments of order r exist using a flat prior if the number of instruments is greater than $m + r$. We discussed the potential of DMC approaches for this case and introduced an extension of DMC that incorporates an acceptance-rejection sampling step within DMC. This Acceptance-Rejection within DMC (ARDMC) method has as attractive property that the generated random drawings are independent, which greatly helps the fast convergence of simulation results, and which facilitates the evaluation of the numerical accuracy. For several cases of simulated and empirical data sets ARDMC outperforms the Gibbs sampler in terms of numerical accuracy.

We leave the following issues as topics for future research. First, the speed of ARDMC can be easily further improved by making use of parallelized computation using multiple core machines and computer clusters, which is less straightforward for MCMC methods. This could reduce the computing time by a substantial factor. Second, one may focus on the Errors in Variables (EV) model and the Simultaneous Equations Model (SEM). Third, as an alternative to a choice between the linear model and IV model, one may use Bayesian Model Averaging (BMA) of the posteriors in the linear and IV model, based on either the marginal or predictive likelihoods of the models, see Zellner et al. (2011). Fourth, the ARDMC procedure may be used to simulate candidate draws for Importance Sampling or the independence chain Metropolis-Hastings algorithm, in cases where one specifies an informative prior for the parameters Π or Σ.

References

Acemoglu, D., Johnson, S., Robinson, J. A., 2001. The colonial origins of comparative development: An empirical investigation. American Economic Review 91 (5), 1369–1401.

Ando, T., Zellner, A., 2010. Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques. Bayesian Analysis 5 (1), 65–96.

Angrist, J. D., Krueger, A. B., 1991. Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics 106 (4), 979–1014.

Bauwens, L., Van Dijk, H. K., 1990. Bayesian limited information analysis revisited. In: Gabszewicz, J. J., Richard, J. F., Wolsey, L. A. (Eds.), Economic Decision-Making: Games, Econometrics and Optimisation: Contributions in Honour of Jacques H. Drèze. North-Holland, Ch. 18, pp. 385–424.

Conley, T. G., Hansen, C. B., McCulloch, R. E., Rossi, P. E., 2008. A semi-

parametric Bayesian approach to the instrumental variable problem. Journal of Econometrics 144 (1), 276–305.

De Pooter, M., Ravazzolo, F., Segers, R., Van Dijk, H. K., 2008. Bayesian near-boundary analysis in basic macroeconomic time series models. In: Chib, S., Griffiths, W., Koop, G., Terrell, D. (Eds.), Bayesian Econometrics. Vol. 23 of Advances in Econometrics. Emerald Group, pp. 331–402.

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

Drèze, J. H., 1977. Bayesian regression analysis using poly-t densities. Journal of Econometrics 6 (3), 329–354.

Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. Wiley, New York.

Geyer, C. J., 1992. Practical Markov Chain Monte Carlo. Statistical Science 7 (4), 473–483.

Hood, W. M. C., Koopmans, T. C. (Eds.), 1950. Studies in Econometric Method. No. 14 in Cowles Commision Monographs. Wiley, New York.

Hoogerheide, L. F., 2006. Essays on Neural Network Sampling methods and Instrumental Variables. PhD thesis (nr. 379 of the Tinbergen Institute Research Series), Erasmus University Rotterdam.

Hoogerheide, L. F., Block, J. H., Thurik, A. R., 2012a. Family background variables as instruments for education in income regressions: A Bayesian analysis. Economics of Education Review 31 (5), 515–523.

Hoogerheide, L. F., Kaashoek, J. F., Van Dijk, H. K., 2007. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks. Journal of Econometrics 139 (1), 154–180.

Hoogerheide, L. F., Opschoor, A., Van Dijk, H. K., 2012b. A class of adaptive Importance Sampling weighted EM algorithms for efficient and robust posterior and predictive simulation. Journal of Econometrics forthcoming.

Kleibergen, F., Van Dijk, H. K., 1998. Bayesian simultaneous equations analysis using reduced rank structures. Econometric Theory 14, 701–743.

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

Koopmans, T. C. (Ed.), 1950. Statistical inference in dynamic economic models. No. 10 in Cowles Commision Monographs. Wiley, New York.

Liu, J., 2001. Monte Carlo strategies in scientific computing. Springer, New York.

Liu, J. S., 1996. Metropolized independent sampling with comparison to rejection sampling and importance sampling. Statistics and Computing 6, 113–119.

Van Dijk, H. K., 2003. On Bayesian structural inference in a simultaneous equation model. In: Stigum, B. P. (Ed.), Econometrics and the philosophy of economics. Vol. 23. Princeton University Press, pp. 642–682.

Wagner, G. G., Burkhauser, R. V., Behringer, F., 1993. The English language public use file of the German socio-economic panel study. The Journal of Human Resources 28 (2), 429–433.

Wagner, G. G., Frick, J. R., Schupp, J., 2007. The German socio-economic panel study (soep) – scope, evolution and enhancements. Schmollers Jahrbuch 127 (1),

139–169.
Zellner, A., 1971. An introduction to Bayesian inference in econometrics. Wiley, New York.
Zellner, A., Ando, T., 2008. A direct Monte Carlo approach for Bayesian analysis of the simultaneous equation model. Working paper.
Zellner, A., Ando, T., 2010a. Bayesian and non-Bayesian analysis of the seemingly unrelated regression model with student-t errors, and its application for forecasting. International Journal of Forecasting 26 (2), 413–434.
Zellner, A., Ando, T., 2010b. A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model. Journal of Econometrics 159 (1), 33–45.
Zellner, A., Ando, T., Baştürk, N., Hoogerheide, L., Van Dijk, H. K., 2011. Direct and indirect Monte Carlo for simultaneous equations, instrumental variables and errors in variables models: On the connection between model structures, data information and efficient posterior simulators. Tinbergen Institute Discussion Paper 2011-137/4.
Zellner, A., Bauwens, L., Van Dijk, H. K., 1988. Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo integration. Journal of Econometrics 38 (1-2), 39–72.

A IV model with m possibly endogenous regressors under a flat prior: derivations of conditional and marginal posterior distributions

This Appendix provides a concise derivation of the conditional and marginal posterior distributions, and the results on properness and posterior moments, that are considered in Section 2.

For the **conditional posterior density of β given Π and Σ**, we use the fact that only u is a function of parameter β in (5), and properties of the multivariate normal distribution. We have $u|V,\Sigma \sim N(\mu_{u|V,\Sigma},\, \omega_{u|V,\Sigma}\, I_T)$. Hence the conditional posterior of β given Π and Σ is:

$$p(\beta \mid \Pi, \Sigma, y, X, Z) \propto p(u|V,\Sigma) \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\omega_{u|V,\Sigma}^{-1}\left(y-\mu_{u|V,\Sigma}-X\beta\right)'\left(y-\mu_{u|V,\Sigma}-X\beta\right)\right)\right\}. \tag{A.1}$$

Completing the sum of squares in (A.1) shows that the conditional posterior of $\beta \mid \Pi, \Sigma$ is the multivariate normal density $N(\mu_{\beta|\Pi,\Sigma}, \Omega_{\beta|\Pi,\Sigma})$, where $\mu_{\beta|\Pi,\Sigma} \equiv (X'X)^{-1}X'(y-\mu_{u|V,\Sigma})$ and $\Omega_{\beta|\Pi,\Sigma} \equiv \omega_{u|V,\Sigma}\,(X'X)^{-1}$.

For the **conditional posterior density of Π given β and Σ**, we use the fact that only V is a function of parameter Π in (5), and properties of multivariate normal distribution. We have $\mathrm{vec}(V)|u,\Sigma \sim N(\mathrm{vec}(\mu_{V|u,\Sigma}), \Omega_{V|u,\Sigma} \otimes I_T)$. Hence the

conditional posterior of Π is:

$$p(\Pi \mid \beta, \Sigma, y, X, Z) \propto p(V|u, \Sigma) \propto \exp\left\{-\frac{1}{2}\text{tr}\left(\Omega_{V|u,\,\Sigma}^{-1}\left(X - \mu_{V|u,\,\Sigma} - Z\Pi\right)'\left(X - \mu_{V|u,\,\Sigma} - Z\Pi\right)\right)\right\}. \tag{A.2}$$

Completing the squares in (A.2) shows that the conditional posterior of $\Pi|\beta, \Sigma$ is the matrix normal distribution $N_{\text{matrix}}(\mu_{\Pi|\beta,\,\Sigma}, \Omega_{V|u,\,\Sigma}, (Z'Z)^{-1})$ with $\mu_{\Pi|\beta,\,\Sigma} \equiv (Z'Z)^{-1}Z'(X - \mu_{V|u,\,\Sigma})$.

The **marginal posterior of β and Π** is obtained by the Inverse-Wishart step on Σ:

$$p(\beta, \Pi \mid y, X, Z) \propto \int_{\Sigma} |\Sigma|^{-(T+m+2)/2} \exp\left\{-\frac{1}{2}\text{tr}\left((u\ V)'(u\ V)\,\Sigma^{-1}\right)\right\} \mathrm{d}\Sigma, \tag{A.3}$$

where the right hand side is the Inverse-Wishart density apart from an integrating constant and the factor $|(u\ V)'(u\ V)|^{T/2}$, so

$$p(\beta, \Pi \mid y, X, Z) \propto \left|(u\ V)'(u\ V)\right|^{-T/2}. \tag{A.4}$$

The **conditional posterior of β given Π** is obtained by using the following determinant decomposition:

$$p(\Pi, \beta \mid y, X, Z) \propto \left|(u\ V)'(u\ V)\right|^{-T/2} = |V'V|^{-T/2}\,(u'M_V u)^{-T/2} \tag{A.5}$$

Completing the squares on β yields:

$$p(\beta, \Pi \mid y, X, Z) \propto |V'V|^{-\frac{T}{2}}\left((T-m)\,s_{\hat{\beta}}^2\right)^{-\frac{T}{2}}\left(1 + \frac{\left(\beta - \hat{\beta}\right)'(X'M_V X)\left(\beta - \hat{\beta}\right)}{(T-m)\,s_{\hat{\beta}}^2}\right)^{-\frac{T}{2}}, \tag{A.6}$$

so that $p(\beta \mid \Pi, y, X, Z)$ is a multivariate t density with location vector $\hat{\beta}$ and scale matrix $s_{\hat{\beta}}^2(X'M_V X)^{-1}$ and $T - m$ degrees of freedom, *given that* $(X'M_V X)$ has full rank m. The latter holds if

$$|X'M_V X| = |V'M_X V|\frac{|X'X|}{|V'V|} > 0 \Leftrightarrow |\Pi'Z'M_X Z\Pi|\frac{|X'X|}{|V'V|} > 0 \tag{A.7}$$

where we have used that $M_X V = M_X(X - Z\Pi) = -M_X Z\Pi$. (A.7) holds if $\Pi'Z'M_X Z\Pi$ has full rank m, which is true if and only if Π has full column rank m.

In a similar fashion, we derive the **conditional posterior of Π given β**:

$$p(\beta, \Pi \mid y, X, Z) \propto (u'u)^{-T/2}\left|S_{\hat{\Pi}}\left(I_m + (S_{\hat{\Pi}})^{-1}\left(\Pi - \hat{\Pi}\right)'Z'M_u Z\left(\Pi - \hat{\Pi}\right)\right)\right|^{-T/2}. \tag{A.8}$$

That is, $p(\Pi \mid \beta, y, X, Z)$ is a matrix t density with location matrix $\hat{\Pi}$, scale matrices $(Z'M_uZ)^{-1}$ and $S_{\hat{\Pi}}$, and $T-k-m+1$ degrees of freedom for any number of endogenous variables m, any number of instruments k and for every value of β.

The **marginal posterior of** β is derived by integrating (A.8):

$$p(\beta \mid y, X, Z) \propto (u'u)^{-T/2} \int_{\Pi} |S_{\hat{\Pi}}|^{-T/2} \left| \left(I_m + (S_{\hat{\Pi}})^{-1} \left(\Pi - \hat{\Pi}\right)' Z'M_uZ \left(\Pi - \hat{\Pi}\right) \right) \right|^{-T/2} \mathrm{d}\Pi, \tag{A.9}$$

$$= (u'u)^{-T/2} |S_{\hat{\Pi}}|^{-(T-k)/2} |Z'M_uZ|^{-m/2} \int_{\Pi} |Z'M_uZ|^{m/2} |S_{\hat{\Pi}}|^{-k/2} \left(I_m + (S_{\hat{\Pi}})^{-1} \left(\Pi - \hat{\Pi}\right)' Z'M_uZ \left(\Pi - \hat{\Pi}\right) \right) \mathrm{d}\Pi, \tag{A.10}$$

$$\propto (u'u)^{-T/2} |S_{\hat{\Pi}}|^{-(T-k)/2} |Z'M_uZ|^{-m/2}, \tag{A.11}$$

where the integral in (A.10) is the matrix t density. The marginal posterior of β in (A.11) is simplified as follows. The third factor (A.11) is:

$$|Z'M_uZ| = (u'M_Zu) \frac{|Z'Z|}{(u'u)} \propto \frac{u'M_Zu}{u'u}. \tag{A.12}$$

The second factor in (A.11) is:

$$|S_{\hat{\Pi}}| = \left| (M_uX)' M_{M_uZ} (M_uX) \right| = |(M_uZ)'M_{M_uX}M_uZ| \frac{|X'M_uX|}{|Z'M_uZ|}, \tag{A.13}$$

where the first factor on the right-hand side of (A.13), the sample covariance matrix (multiplied by $T-m-1$) of the residuals in a regression of Z on X and u, is equal to

$$(M_uZ)'M_{M_uX}M_uZ = (M_XZ)'M_{M_Xu}M_XZ = (M_XZ)'M_{M_Xy}M_XZ, \tag{A.14}$$

which does not depend on β; in (A.14) we used $M_Xu = M_X(y - X\beta) = M_Xy$. Therefore

$$|\Sigma_{\hat{\Pi}}| \propto \frac{|X'M_uX|}{|Z'M_uZ|} \propto \frac{u'M_Xu\,|X'X|}{u'u} \left(\frac{u'M_Zu\,|Z'Z|}{u'u} \right)^{-1} \propto (u'M_Zu)^{-1}. \tag{A.15}$$

Substituting (A.12) and (A.15) into (A.11) yields:

$$p(\beta \mid y, X, Z) \propto (u'u)^{-\frac{T-m}{2}} (u'M_Zu)^{\frac{T-k-m}{2}}, \tag{A.16}$$

which is a t-density multiplied by a polynomial, or

$$p(\beta \mid y, X, Z) \propto \left(\frac{u'M_Zu}{u'u} \right)^{\frac{T-k-m}{2}} (u'u)^{-\frac{k}{2}}, \tag{A.17}$$

where the ratio $\frac{u'M_Zu}{u'u} < 1$ for any β; for 'large enough' β (i.e., $||\beta||$ large enough) the ratio $\frac{u'M_Zu}{u'u}$ is bounded from below and above by ratios of positive eigenvalues of $X'M_ZX$ and $X'X$. Therefore, the tail behavior and properness of $p(\beta \mid y, X, Z)$ are determined by the factor $(u'u)^{-\frac{k}{2}}$, which is an m-dimensional t density with $r = k - m$ integer degrees of freedom. Therefore, $p(\beta \mid y, X, Z)$ is an improper density for $k \leq m$ (exact or under-identification), and a proper density for $k > m$ (over-identification).

The **marginal posterior of** Π is derived by integrating (A.6):

$$p\left(\Pi \mid y, X, Z\right) \propto |V'V|^{-\frac{T}{2}} \left((T-m)\, s_{\hat{\beta}}^2\right)^{-\frac{T}{2}} \left|\frac{X'M_VX}{s_{\hat{\beta}}^2}\right|^{-1/2} \times \int_{\beta} \left|\frac{X'M_VX}{s_{\hat{\beta}}^2}\right|^{1/2} \left(1 + \frac{\left(\beta - \hat{\beta}\right)' \left(X'M_VX\right) \left(\beta - \hat{\beta}\right)}{(T-m)\, s_{\hat{\beta}}^2}\right)^{-\frac{T}{2}} \mathrm{d}\beta, \tag{A.18}$$

$$\propto |V'V|^{-\frac{T}{2}} \left(s_{\hat{\beta}}^2\right)^{-\frac{T}{2}} \left|\frac{X'M_VX}{s_{\hat{\beta}}^2}\right|^{-1/2} \propto |V'V|^{-\frac{T}{2}} \left(s_{\hat{\beta}}^2\right)^{-\frac{T-m}{2}} |X'M_VX|^{-1/2}, \tag{A.19}$$

where the integrand in (A.18) is a multivariate t density. Inserting

$$|X'M_VX| = |V'M_XV| \frac{|X'X|}{|V'V|} \tag{A.20}$$

$$\sigma_{\hat{\beta}}^2 \propto |\left(M_Xy\right)' M_{M_VX} \left(M_Xy\right)| = |\left(M_XV\right)' M_{M_yX} \left(M_Xv\right)| \frac{(y'M_Xy)}{|V'M_XV|} \propto \frac{|V'M_{(y\ X)}V|}{|V'M_XV|} \tag{A.21}$$

into (A.19) yields:

$$p\left(\Pi \mid y, X, Z\right) \propto |V'V|^{-T/2} \left(\frac{|V'M_XV|}{|V'V|}\right)^{-1/2} \left|V'M_{(y\ X)}V\right|^{\frac{T-m}{2}} |V'M_XV|^{-\frac{T-m}{2}}. \tag{A.22}$$

Substituting

$$M_XV = M_X\left(X - Z\Pi\right) = M_XZ\Pi \tag{A.23}$$

$$M_{(y\ X)}v = M_{(y\ X)}\left(X - Z\Pi\right) = M_{(y\ X)}Z\Pi. \tag{A.24}$$

into (A.22) yields:

$$p\left(\Pi \mid y, X, Z\right) \propto |V'V|^{-\frac{T-1}{2}} |\Pi'Z'M_XZ\Pi|^{\frac{T-m-1}{2}} \left|\Pi'Z'M_{(y\ X)}Z\Pi\right|^{-\frac{T-m}{2}}, \tag{A.25}$$

a matrix t-density multiplied by a rational function, or

$$p\left(\Pi \mid y, X, Z\right) \propto |V'V|^{-\frac{T-1}{2}} \left(\frac{|\Pi' Z' M_X Z \Pi|}{\left|\Pi' Z' M_{(y\ X)} Z \Pi\right|} \right)^{\frac{T-m}{2}} |\Pi' Z' M_X Z \Pi|^{-\frac{1}{2}}, \qquad \text{(A.26)}$$

where $|V'V|^{-\frac{T-1}{2}}$ is a density kernel of a proper matrix t distribution of which the first few moments are finite (given that T is not very small), and the ratio $\frac{|\Pi' Z' M_X Z \Pi|}{\left|\Pi' Z' M_{(y\ X)} Z \Pi\right|}$ is bounded from below and above by ratios of positive eigenvalues of $Z' M_X Z$ and $Z' M_{(y\ X)} Z$. So, the properness of $p\left(\Pi \mid y, X, Z\right)$ is determined by the factor $|\Pi' Z' M_X Z \Pi|^{-\frac{1}{2}}$, which is integrable if and only if $k > m$ (over-identification). In the latter case, the first few moments – i.e., at least up to the fourth moment – exist (given that T is not very small).