CORE DISCUSSION PAPER
2005/29

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

Lennart F. HOOGERHEIDE[1], Johan F. KAASHOEK[2] and Herman K. VAN DIJK[3]

March 2005

Abstract

Likelihoods and posteriors of instrumental variable regression models with strong endogeneity and/or weak instruments may exhibit rather non-elliptical contours in the parameter space. This may seriously affect inference based on Bayesian credible sets. When approximating such contours using Monte Carlo integration methods like importance sampling or Markov chain Monte Carlo procedures the speed of the algorithm and the quality of the results greatly depend on the choice of the importance or candidate density. Such a density has to be 'close' to the target density in order to yield accurate results with numerically efficient sampling. For this purpose we introduce neural networks which seem to be natural importance or candidate densities, as they have a universal approximation property and are easy to sample from. A key step in the proposed class of methods is the construction of a neural network that approximates the target density accurately. The methods are tested on a set of illustrative models. The results indicate the feasibility of the neural network approach.

Keywords: instrumental variables, reduced rank, importance sampling, Markov chain Monte Carlo, neural networks, Bayesian inference, credible sets. **JEL classification:** C11, C15, C45

[1]Tinbergen and Econometric Institutes, Erasmus University Rotterdam. E-mail: lhoogerheide@few.eur.nl

[2]Econometric Institute, Erasmus University Rotterdam. E-mail: kaashoek@few.eur.nl

[3]Econometric and Tinbergen Institutes, Erasmus University Rotterdam. E-mail: hkvandijk@few.eur.nl. Part of this paper was written when the third author was visiting scholar at CORE, Université catholique de Louvain

A preliminary version of this paper (Hoogerheide, Kaashoek and Van Dijk (2004)) was presented at the 2003 EC^2 meeting in London. The authors are indebted to Andrew Chesher, Geert Dhaene and two anonymous referees for helpful comments which have led to substantial improvements. All remaining errors are the authors' responsibility. This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

1 Introduction

There exist classes of statistical and econometric models where the conditional distributions of parameters of interest have known analytical properties such that one may construct regular Bayesian credible sets which are elliptically shaped. As a consequence, conditional Bayesian inference may be performed in a standard way. However, the joint and marginal distributions of the parameters have no known analytical properties nor elliptically shaped credible sets. So, it is not trivial to perform inference on the joint distribution. This may have strong effects on the measurement of uncertainty of forecasts and of certain policy measures. For instance, in labor market models it is important to know whether a certain credible set of the policy effects of training programs has a strongly asymmetric shape. In models of international financial markets, used for hedging currency risk, knowledge of a strongly non-elliptical credible set is important for the specification of an optimal hedging decision under risk. For details on such models we refer to *e.g.* Angrist and Imbens (1994) and Bos, Mahieu and Van Dijk (2000) and the references cited there.

A second issue is that one may have great difficulties trying to simulate random drawings from such a class of non-elliptical joint distributions. This feature is useful for inference on such nonlinear functions of parameters of interest as impulse responses, see Strachan and Van Dijk (2004). Even if it is relatively easy to simulate (pseudo-) random drawings from the conditional distributions, multi-modality and/or high correlations may cause the Gibbs sampler to be extremely inefficient or even yield erroneous results.

A canonical case is the example given by Gelman and Meng (1991), where the conditional distributions are known to be normal and thus it is easy to perform conditional inference and it is easy to simulate random drawings from the conditional distributions; however the joint density is not known in terms of analytical properties. This class of conditionally normal distributions contains bimodal joint distributions that are not trivial to sample from, as the Gibbs sampler may seriously suffer from the fact that the two modes are far apart.

A first contribution of this paper is that we extend this analysis to the case of linear models with reduced rank. We focus on the class of instrumental variable (IV) regression models with possibly endogenous regressors. Traditionally, these models are used as a special case of structural equation systems. More recently, these models are applied to uncover local average treatment effects, see Angrist, Imbens and Rubin (1996). Under certain weak priors the conditional posterior distributions in the IV regression model are Student t, that is, at least if they are proper. This class of models may exhibit reduced rank of the parameter matrix which may be due to varying degrees of identification, endogeneity and quality of instruments.[1] In the presence of weak instruments the posterior may display highly non-elliptical contours.

When approximating such non-standard contours using Monte Carlo integration methods like importance sampling or Markov chain Monte Carlo procedures[2] the speed of the algorithm and the quality of the results greatly depends on the choice of the importance or candidate density. Such a density has to be 'close' to the target density in order to yield accurate results with

[1] We note that reduced rank occurs also in cointegration models where one determines the number of stable economic relations; and in factor models where the number of common factors needs to be determined; or in errors in variables models. A more detailed analysis is in progress and will be reported in a later paper.

[2] The theory of Markov chain Monte Carlo (MCMC) methods starts with Metropolis et al. (1953) and Hastings (1970). An important technical paper on MCMC methods is due to Tierney (1994). Well-known econometric studies are provided by Chib and Greenberg (1996) and Geweke (1999). Indirect independence sampling methods such as importance sampling (IS) have also been successfully applied within Bayesian inference. Importance sampling, see Hammersley and Handscomb (1964), has been introduced in Bayesian inference by Kloek and Van Dijk (1978) and is further developed by Van Dijk and Kloek (1980,1984) and Geweke (1989).

numerically efficient sampling.

A second contribution of this paper is that we introduce a class of neural network sampling methods which allow for sampling from a target (posterior) distribution that may be multi-modal or skew, or exhibit strong correlation among the parameters. That is, a class of methods to sample from non-elliptical distributions.

Neural network sampling algorithms consist of two main steps, which are summarized as follows. In the first step a neural network is constructed that approximates the target density reasonably well. In the second step this neural network is embedded in a Metropolis-Hastings (MH) or importance sampling (IS) algorithm.

With respect to the first step we emphasize that an important advantage of neural network functions is their 'universal approximation property'. That is, neural network functions can provide approximations of any square integrable function to any desired accuracy, see Gallant and White (1989). As an application of Kolmogorov's general superposition theorem, the neural network approximation property is further explored by Hecht-Nielsen (1987). Proofs concerning neural network approximations for specific configurations can be found in Gallant and White (1989), Hornik et al. (1989), and Leshno et al. (1993). Stinchcombe (1988,1989) shows that it is the presence of intermediate layers with sufficiently many parallel processing elements that is essential for feedforward networks to possess universal approximation capabilities, and that sigmoid activation functions are not necessary for universal approximation. This approximation property implies that the algorithm can handle certain non-elliptical target distributions, like multi-modal, extremely skew, strongly correlated or fat-tailed distributions.

In the second step this neural network is used as an importance function in IS or as a candidate density in MH. In a 'standard' case of Monte Carlo integration, the MH candidate density function or the importance function is uni-modal. If the target (posterior) distribution is multi-modal then a second mode may be completely missed in the MH approach and some drawings may have huge weights in the IS approach. As a consequence the convergence behavior of these Monte Carlo integration methods is rather uncertain. Thus, an important problem is the choice of the candidate or importance density especially when little is known *a priori* about the shape of the target density. Given a reasonably accurate approximation of the neural network constructed in the first step, an important advantage is that neural networks are relatively easy to sample from. This depends, of course, on the specification of the network.

The proposed methods are applied on a set of illustrative examples of conditionally normal distributions and posterior distributions in an instrumental variable regression model. Our results indicate that the neural network approach is feasible in cases where a 'standard' MH, IS or Gibbs approach would fail or be rather slow.[3]

The outline of the paper is as follows. In section 2 we consider the shape of posterior densities in a simple IV regression model for simulated data; it is shown that the shapes of Bayesian credible sets depend on the quality of instruments and the level of endogeneity. In section 3 we discuss how to construct a neural network approximation to a density, how to sample from a neural network density, and how to use these drawings within the IS or MH algorithm. Section 4 shows the feasibility of our approach in a simple example of a bivariate conditionally normal distribution. Section 5 illustrates our algorithms in examples of IV regressions with simulated data. Conclusions are given in section 6 and some derivations are given in the appendix.

[3]We are indebted to two anonymous referees who suggested to make use of more sophisticated Monte Carlo methods like bridge sampling and to use other flexible approximating densities involving Hermite polynomials. This is an area of further research as we indicate in our conclusions

2 On the shape of posterior densities and Bayesian credible sets in instrumental variable regression models with several degrees of identification, endogeneity and instrument quality

As we mentioned in the introduction, there exist several models in which the conditional posterior distributions of parameters of interest have known properties but the joint does not. In this section we consider a class of such models, instrumental variable (IV) regression models with possibly endogenous regressors.

First, we give an example of a well-known IV regression. Consider the stylized wage regression popular in empirical labor studies:

$$y_1 = \beta\, y_2 + \gamma\, x_1 + u_1, \tag{1}$$

where y_1 is the log of hourly wage, y_2 denotes education and x_1 captures work experience – all in deviations from their mean values. The structural parameter of interest is β, the rate of return to schooling. However, in order to make inference on β, one should take into account that y_2 is possibly endogenous: y_2 and u_1 may be highly correlated owing to the omission of a variable measuring (unobservable) ability, which is expected to be highly correlated with education. The problem is that potential instruments for y_2 are hard to find as these variables must be correlated with education but uncorrelated with unobserved ability. Angrist and Krueger (1991) suggest using quarter of birth as a dummy variable, as this seems uncorrelated with ability and affects years of schooling weakly, through a combination of the age at which a person begins school and the compulsory education laws in a person's state. Staiger and Stock (1997) show that inference on the rate of return to schooling can be greatly affected by the weak quarter of birth instruments.

In the sequel of this section we consider the joint, conditional and marginal posterior distributions of the parameters in a simple IV regression model with a weak prior. We show how the shapes of the posterior distributions depend on the varying degrees of identification, endogeneity and quality of instruments. In section 5 we use our neural network sampling methods to generate random drawings from some of the joint densities that are shown in this section.

The model

We consider the following possibly overidentified Instrumental Variables (IV) model, which is also known as the incomplete simultaneous equations model (INSEM). Following Zellner, Bauwens and Van Dijk (1988), let:

$$y_1 = y_2\beta + \varepsilon \tag{2}$$
$$y_2 = X\pi + v \tag{3}$$

where y_1 is a $(T \times 1)$ vector of observations on the endogenous variable that is to be explained, y_2 is a $(T \times 1)$ vector of observations on the explanatory endogenous variable, X is a $(T \times k)$ matrix of weakly exogenous variables; β is a scalar structural parameter of interest, π is a $(k \times 1)$ vector of reduced form parameters. We further assume that the rows of the matrix of error terms $U \equiv (\varepsilon\ v)$ are independently normally distributed with the (2×2) covariance matrix Σ with elements σ_{ij} $(i, j = 1, 2)$. We note that (2)-(3) may be further interpreted as an errors in variables model, see *e.g.* Zellner, Bauwens and Van Dijk (1988).

In the derivations we use the following notation: the symbols ε, v and U denote $\varepsilon = y_1 - y_2\beta$, $v = y_2 - X\pi$ and $U = (\varepsilon\ v) = (y_1 - y_2\beta\ \ y_2 - X\pi)$, i.e. functions of the parameters β and π (and the data y_1, y_2, X) instead of the real error terms. The matrix M_A denotes the $T \times T$ projection matrix $M_A \equiv I - A(A'A)^{-1}A'$.

The joint posterior of (β, π)

A kernel of the likelihood function is given by:

$$L(\beta, \pi, \Sigma | y_1, y_2, X) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}U'U)\right]. \tag{4}$$

We specify the following non-informative prior density:

$$p(\beta, \pi, \Sigma) \propto |\Sigma|^{-h/2} \text{ with } h > 0. \tag{5}$$

From the likelihood (4) and the prior (5) we obtain the joint posterior density of the parameters (β, π, Σ):

$$p(\beta, \pi, \Sigma | y_1, y_2, X) \propto |\Sigma|^{-(T+h)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}U'U)\right]. \tag{6}$$

Using properties of the inverted Wishart distribution (see Zellner (1971) and Bauwens and Van Dijk (1990)), Σ is integrated out of the joint posterior in (6), resulting in the joint posterior for (β,π):

$$p(\beta, \pi | y_1, y_2, X) \propto |U'U|^{-(T+h-3)/2}. \tag{7}$$

Choosing $h = 3$ in the prior density kernel (5) leads to the following joint posterior of (β, π):

$$p(\beta, \pi | y_1, y_2, X) \quad \propto \quad |U'U|^{-T/2} = |(\varepsilon\ v)'(\varepsilon\ v)|^{-T/2} \tag{8}$$

$$= \quad \left| \begin{array}{cc} (y_1 - y_2\beta)'(y_1 - y_2\beta) & (y_1 - y_2\beta)'(y_2 - X\pi) \\ (y_2 - X\pi)'(y_1 - y_2\beta) & (y_2 - X\pi)'(y_2 - X\pi) \end{array} \right|^{-T/2} \tag{9}$$

Although this function may seem to be a regular one, there is an asymptote at $\pi = 0$. For $\pi = 0$ the posterior density kernel in (9) reduces to the constant $((y_1'y_1)(y_2'y_2) - (y_1'y_2)^2)^{-T/2}$, so that for $\pi = 0$ the conditional posterior density of β is improper.

Although improper on $\mathbb{R}^{k+1}$, the posterior in (9) can be made proper by restricting β and/or π to a certain area. In that case it depends greatly on the data y_1, y_2 and X, whether the asymptote at $\pi = 0$ still dominates the analysis.

For illustrative purposes, the posterior kernel in (9) is calculated for simulated data sets from (2) - (3) with $k = 1$, $T = 100$, $\beta = 0$, $\sigma_{11} = \sigma_{22} = 1$ for nine cases. Three different cases of identification (or quality of instruments) are considered: non identification/irrelevant instruments ($\pi = 0$); weak identification/weak instruments ($\pi = 0.1$); strong identification/good instruments ($\pi = 1$). These cases are combined with three cases of endogeneity, i.e. three different values of the correlation $\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ between the error terms ε and v: strong ($\rho = 0.99$), medium ($\rho = 0.5$) and no ($\rho = 0$) degree of endogeneity. For the non-identified case with strong endogeneity the simulated data are shown in Figure 1. In this case with $k = 1$ we have a $(T \times 1)$ vector of instruments which we denote by x. Notice the high correlation between y_1 and y_2 and that y_2 and x look like uncorrelated white noise series.

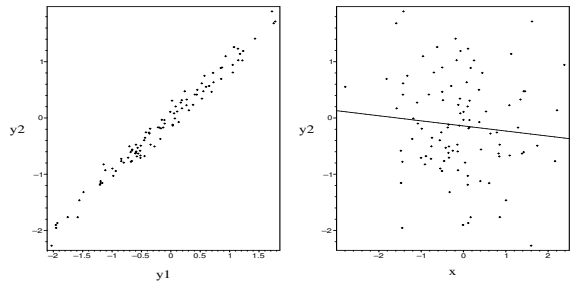

Figure 1: Scatter plots of simulated data in the case of no identification ($\pi = 0$) and strong endogeneity ($\rho = 0.99$)

Figure 2 shows contour plots of the joint posterior kernel of β and π in (9) for our nine simulated data sets. The posterior kernels are normalized over the displayed range. The contour plots reveal that there are three typical shapes of the graph of the joint posterior of β and π: bell-shape, multimodality and elongated ridges.

Note that in the three cases of simulated data sets with strong instruments ($\pi = 1$), the contour plots do not show a ridge at $\pi = 0$. The reason is that the value of the joint posterior kernel for $\pi = 0$ is relatively very small as compared to the value of the joint posterior kernel at its mode $(\tilde{\beta}, \tilde{\pi})$ with $\tilde{\beta} = y_1'x/y_2'x, \tilde{\pi} = y_2'x/x'x$.[4] If we consider the contour plot of the posterior kernel raised to the power 1/20, so that the contour plot also shows the contours for much lower values of the posterior kernel, we observe also in this case of strong identification the presence of multimodality or an elongated ridge around the line $\pi = 0$; see Figure 3. So, even in the presence of good instruments and no/medium endogeneity the contours are, strictly speaking, not elliptical. However, if one restricts the region of integration to a certain bounded area the influence of these tiny ridges on inference is negligible; then one may for practical purposes consider the joint posterior distribution of β and π as elliptical.

[4]The ratio between the posterior kernel in (9) at its mode and its value for $\pi = 0$ is:

$$\left[1 - \frac{r_{y_2,x}^2 + r_{y_1,x}^2 - 2r_{y_1,x}r_{y_2,x}r_{y_1,y_2}}{1 - r_{y_1,y_2}^2}\right]^{-T/2}$$

where $r_{y_2,x} \equiv y_2'x/\sqrt{y_2'y_2\ x'x}$, etc. In our simulation example with $\beta = 0$ we have $r_{y_1,x} \approx 0$, so that the ratio is determined by $r_{y_2,x}^2$ (quality of instrument) and r_{y_1,y_2}^2 (level of endogeneity). The stronger the instruments and the stronger the endogeneity, the smaller the ratio and the (relatively) lower is the ridge at $\pi = 0$. Note that a relatively low ridge at $\pi = 0$ does not immediately imply elliptical contours, see e.g the multimodal posterior in the case of a simulated data set with weak instruments ($\pi = 0.1$) and strong endogeneity ($\rho = 0.99$) in Figure 2.

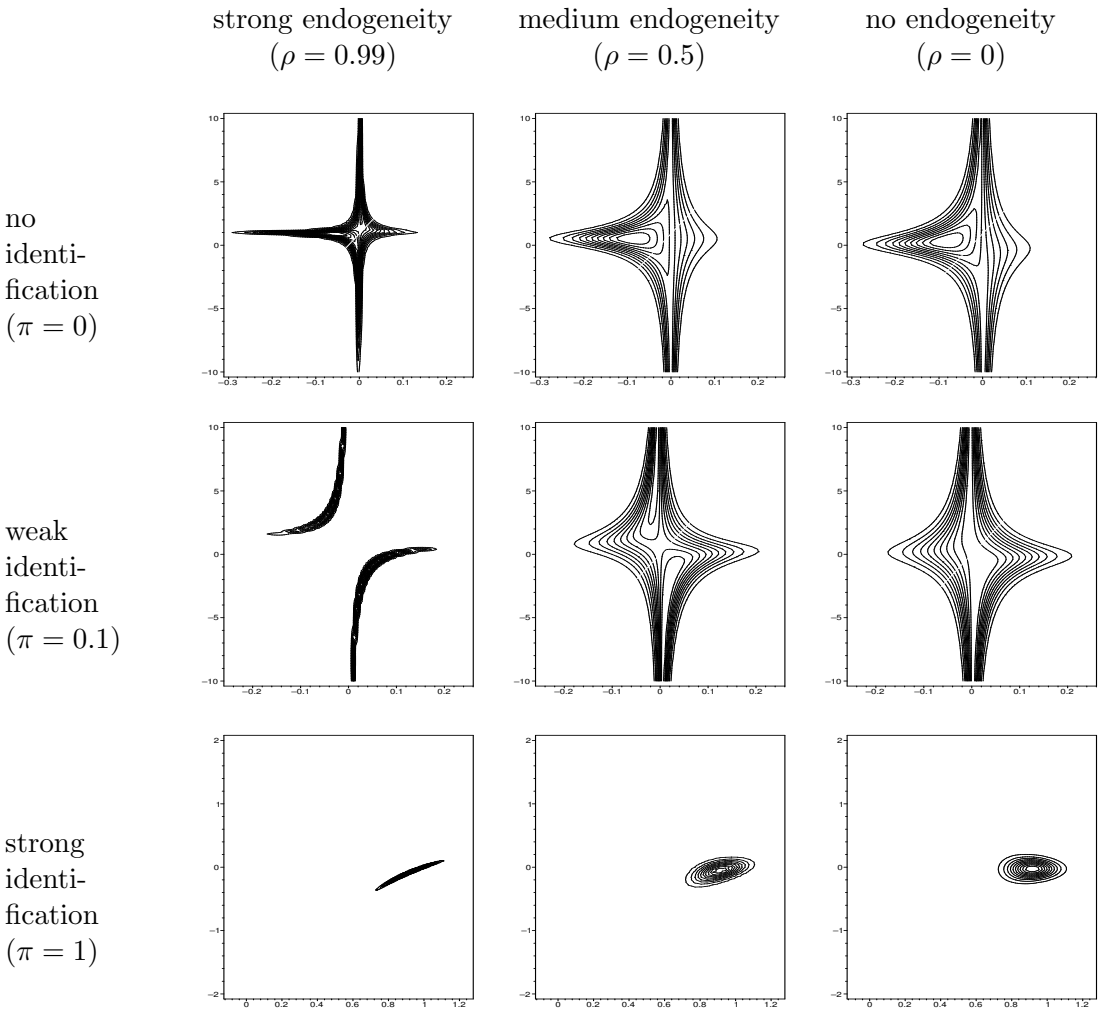

Figure 2: Contour plots in the $\pi \times \beta$ plane: joint posterior kernel of π and β in (9) in IV model for nine simulated data sets; three cases of identification ($\pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity)

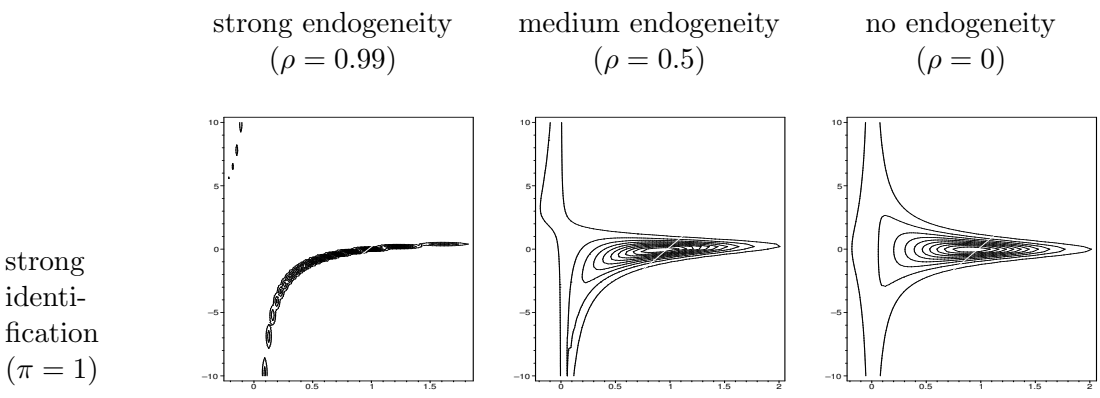

Figure 3: Contour plots in the $\pi \times \beta$ plane: joint posterior kernel of β and π in (9) raised to the power 1/20 in IV model for three simulated data sets; the case of strong identification ($\pi = 1$) combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity)

2.1 Weak and strong structural inference: The conditional and marginal posterior of β

In appendix A the conditional posterior density of β given π and the marginal posterior density of β are derived. We summarize the results in two propositions:

Proposition 1: *In the IV regression model (2)-(3) with prior (5) the conditional posterior density of β given π (with $\pi \neq 0$) is a Student t density with mode $\hat{\beta} \equiv (y_2' M_v y_2)^{-1}(y_2' M_v y_1)$, scale $s^2_{\hat{\beta}}(y_2' M_v y_2)^{-1}$ and $(T-1)$ degrees of freedom:*

$$p(\beta|\pi, y_1, y_2, X) = \frac{c}{\sqrt{s^2_{\hat{\beta}}(y_2' M_v y_2)^{-1}}} \left[1 + \frac{1}{T-1} \frac{(\beta - \hat{\beta})^2}{s^2_{\hat{\beta}}(y_2' M_v y_2)^{-1}}\right]^{-T/2} \tag{10}$$

where $(T-1)s^2_{\hat{\beta}} \equiv (y_1 - y_2\hat{\beta})' M_v (y_1 - y_2\hat{\beta})$ and c is a constant that only depends on T. Moments are finite up to the order $T-1$. For $\pi \to 0$ the conditional posterior variance of β tends to ∞ as in this case $y_2' M_v y_2 \to 0$ (if $\pi = 0$ then $v \equiv y_2 - x\pi = y_2$). For $\pi = 0$ the conditional posterior density of β does not exist or, in other words, is an improper uniform distribution on $(-\infty, \infty)$. For $\pi \neq 0$ HPD regions are elliptical.

Proposition 2 (Drèze (1976, 1977)): *In the IV regression model (2)-(3) with prior (5) the marginal posterior density of β is proportional to the ratio of two Student t kernels:*

$$p(\beta|y_1, y_2, X) \quad \propto \quad \frac{[(y_1 - y_2\beta)'(y_1 - y_2\beta)]^{-(T-1)/2}}{[(y_1 - y_2\beta)' M_X (y_1 - y_2\beta)]^{-(T-k-1)/2}}. \tag{11}$$

This density is known as the 1-1 ratio or poly t density. Structural inference on β depends on the level of identification. Moments exist up to the order of overidentification $(k-1)$.

Corollary 2: *(i) Given a few weak instruments the marginal posterior of β with kernel specified by (11) may be bimodal;*
(ii) Given strong instruments the marginal posterior of β has a bell-shaped graph;
(iii) Many possibly irrelevant instruments give a bell-shaped marginal posterior of β.

So, the marginal posterior density of β tends to a bell shaped function as long as the number of instruments k becomes large enough. This seems to be a paradoxical result: the presence of many (possibly *irrelevant*) instruments gives a bell-shaped function. In other words, even if the *quality* of the instruments is poor, a large *quantity* still yields a bell-shaped marginal posterior of β. This result appeared in an informal way in Maddala (1976), commenting on Drèze (1976).

Figure 4 shows the marginal posterior of β in (11) for our nine simulated data sets. The posterior kernels are normalized over the displayed range. Notice the bimodality in the case of the weak instrument and strong endogeneity. Also note the bell-shape in the cases with a strong instrument.

Figure 5 shows the marginal posterior kernel of β in (11) for the simulated data set corresponding to the case of weak identification and strong endogeneity if independent series of standard Gaussian noise are added to the set of instruments. Clearly the graph of the marginal posterior kernel tends to a bell-shape if many irrelevant instruments are added to the model.

2.2 Impossible restricted reduced form inference: The conditional and marginal posterior of π

In appendix A the conditional posterior density of π given β and the marginal posterior density kernel of π are derived. We summarize the results in two propositions:

Proposition 3: *In the IV regression model (2)-(3) with prior (5) the conditional posterior density of π given β is a Student t density with mode $\hat{\pi} \equiv (X'M_\varepsilon X)^{-1}(X'M_\varepsilon y_2)$, scale $s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1}$ and $(T-k)$ degrees of freedom:*

$$\begin{aligned} p(\pi|\beta, y_1, y_2, X) &= c_2\, |s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1}|^{-1/2} \times \\ &\quad \times \left[1 + \frac{1}{T-k}(\pi-\hat{\pi})'(s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1})^{-1}(\pi-\hat{\pi})\right]^{-T/2} \end{aligned} \tag{12}$$

where $(T-k)s^2_{\hat{\pi}} \equiv (y_2 - X\hat{\pi})'M_\varepsilon(y_2 - X\hat{\pi})$ and c_2 is a scaling constant that only depends on T and k. For all values of β this density exists. Moments are finite up to the order $T-k$. HPD regions are elliptical.

Proposition 4 (Kleibergen and Van Dijk (1994, 1998)): *In the IV regression model (2)-(3) with prior (5) the marginal posterior density of π is proportional to the ratio of a product of two Student t kernels in the numerator and one Student t kernel in the denominator:*

$$p(\pi|y_1, y_2, X) \propto \frac{[(y_2 - X\pi)'(y_2 - X\pi)]^{-(T-1)/2}\ (\pi'X'M_{[y_1\ y_2]}\ X\pi)^{-(T-1)/2}}{(\pi'X'M_{y_2}X\pi)^{-(T-2)/2}} \tag{13}$$

$$\begin{aligned} &= [(y_2 - X\pi)'(y_2 - X\pi)]^{-(T-1)/2} \times \\ &\quad \times (\pi'X'M_{y_2}X\pi)^{-1/2}\left(\frac{\pi'X'M_{y_2}X\pi}{\pi'X'M_{[y_1\ y_2]}\ X\pi}\right)^{(T-1)/2} \end{aligned} \tag{14}$$

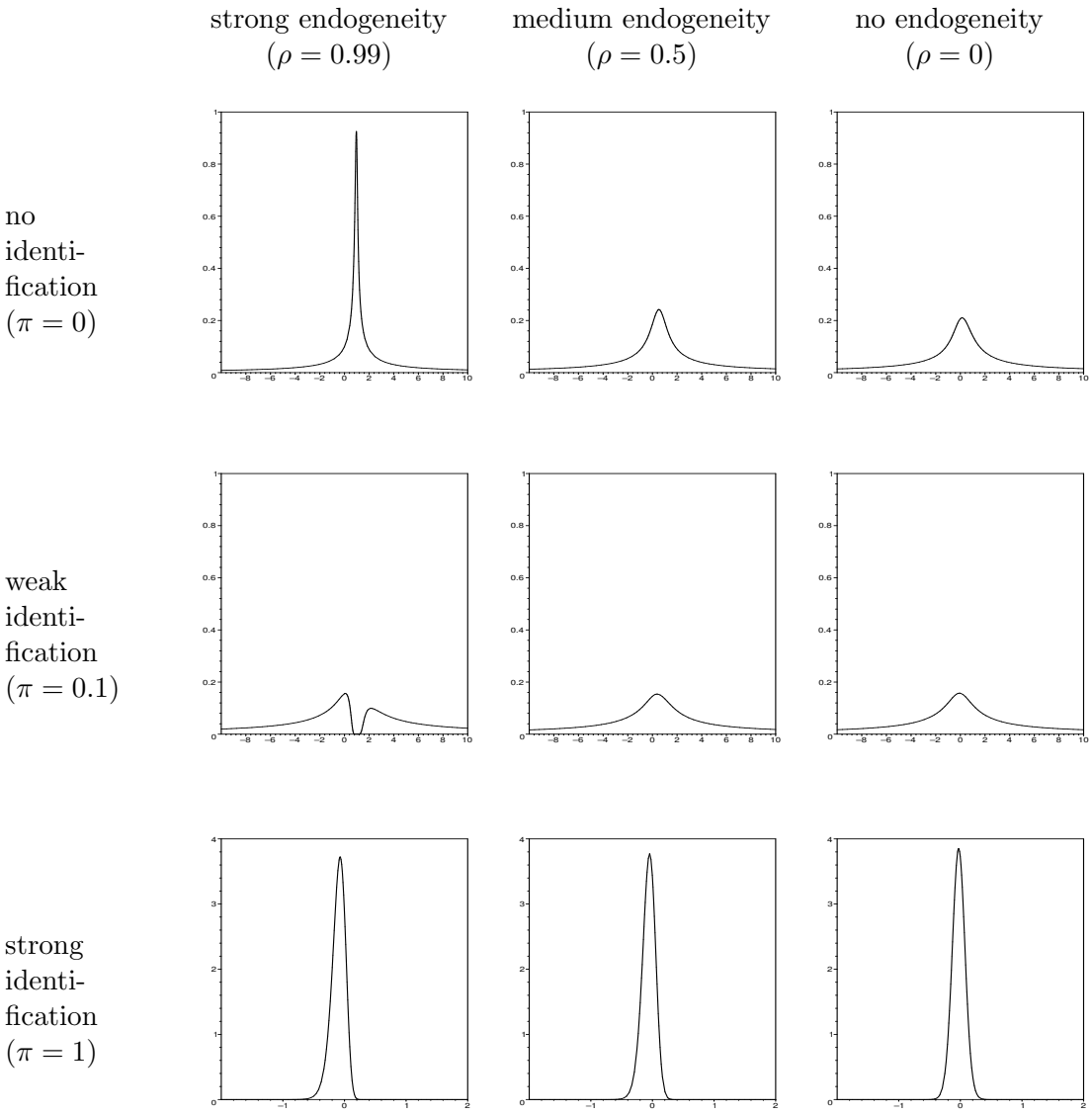

Figure 4: Marginal posterior kernel of β in (11) in IV model for nine simulated data sets; three cases of identification ($\pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity)

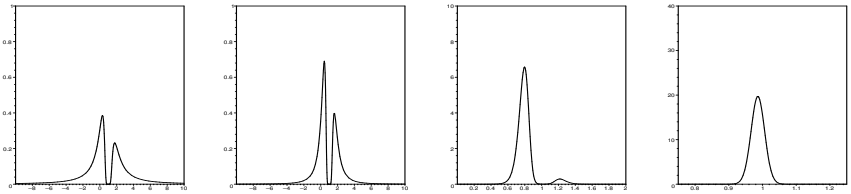

Figure 5: Marginal posterior kernel of β in (11) in IV model for simulated data with weak identification ($\pi = 0.1$) and strong endogeneity ($\rho = 0.99$) after adding 1, 2, 15 or 75 irrelevant (i.i.d. N(0,1)) instruments, respectively

This density is known as the 2-1 poly t density. Reduced form inference on π is not possible, as this is not a proper density. When π tends to zero then an asymptote occurs (because of the term $(\pi' X' M_{y_2} X \pi)^{-1/2}$).

Notice that the result that the marginal posterior of π is not a proper density does not depend on the quality or quantity of the instruments nor on the endogeneity in the data. So, forecasting is not possible when using the restricted reduced form, unless the region of integration of π is truncated, the effect of which is not known a priori. However, it may occur that the data are such that the asymptote will not be noticed in the computations; this may happen if the mode of the joint posterior of (β, π) occurs far away from $\pi = 0$. Figure 6 shows the marginal posterior of π in (14). Notice that each plot reveals an asymptote at $\pi = 0$; however, for the cases of strong identification the spike near $\pi = 0$ is very narrow and relatively far away from the bell-shaped part of the graph.

Only if the restriction that y_2 is not an endogenous regressor is imposed on the model beforehand, i.e. $\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = 0$, we obtain a proper marginal density of π. Specifying the non-informative prior density $p(\beta, \pi, \sigma_{11}, \sigma_{22}) \propto \sigma_{11}^{-1/2}\sigma_{22}^{-1/2}$, and integrating out σ_{11} and σ_{22} using properties of the inverted Gamma distribution (see Zellner (1971)) yields the joint posterior of β and π:

$$p(\beta, \pi | y_1, y_2, X) \propto [(y_1 - y_2\beta)'(y_1 - y_2\beta)]^{-T/2}[(y_2 - X\pi)'(y_2 - X\pi)]^{-T/2} \tag{15}$$

The posterior distributions of β and π are independent Student t with $T - 1$ and $T - k$ degrees of freedom, respectively.

We summarize the results on the joint, conditional and marginal distributions in two tables. Table 1 gives an overview of the possible shapes of the joint posterior kernel of β and π in a simple IV regression model with $k = 1$ instrument for different cases of simulated data.[5] Table 2 gives an overview of the classes of conditional and marginal densities in IV regression models.

[5]We have repeated the experiment with a different seed of the random number generator. In four of the nine cases bimodality only showed up in the contour plot in one of the two simulations; this is denoted with 'possibly bimodality'.

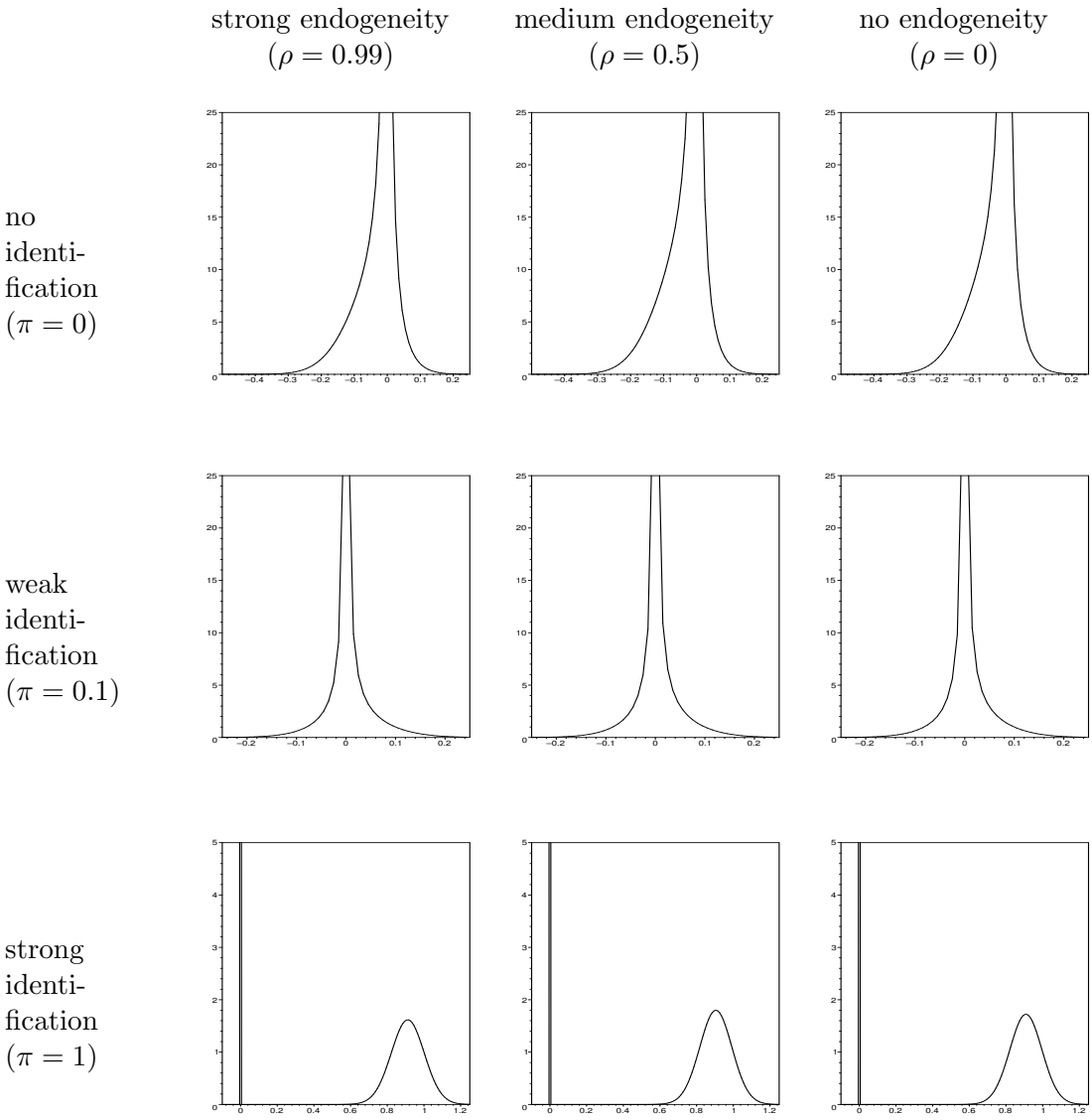

Figure 6: Marginal posterior kernel of π in (13) in IV model for nine simulated data sets; three cases of identification ($\pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity). Note that each figure reveals an asymptote at $\pi = 0$.

Table 1: Shape of the posterior density kernel of β and π in the IV regression model (2)-(3) with one instrument with weak prior (5) for nine situations

		Degree of endogeneity		
		strong	medium	no
Level of identification/ Quality of instruments	no	ridges and possibly bimodality	ridges and possibly bimodality	ridges
	weak	ridges and bimodality	ridges and possibly bimodality	ridges and possibly bimodality
	strong	nearly elliptical	elliptical	elliptical

Table 2: Classes of posterior densities in IV models with weak prior (5)

	Conditional density	Marginal density
β	Student t, improper for $\pi = 0$	1-1 poly t (Drèze (1976)), i.e. ratio of two Student t kernels: improper for small k
π	Student t, proper for all β	2-1 poly t (Kleibergen and Van Dijk (1994,1998)), i.e. ratio of product of two Student t kernels in numerator and one Student t kernel in denominator: improper

3 Approximating with and sampling from neural networks

Consider a certain distribution, for example a posterior distribution, with density kernel $p(\theta)$ with $\theta \in \mathbb{R}^n$. In the case of the IV regression model in the previous section we considered $\theta = (\beta, \pi')'$. Suppose the aim is to investigate some of the characteristics of $p(\theta)$, for example the mean and/or covariance matrix of a random vector $\theta \sim p(\theta)$. The approach followed in this paper consists of the following steps:

1. Find a neural network approximation $nn : \mathbb{R}^n \to \mathbb{R}$ to the target density kernel $p(\theta)$.
2. Obtain a sample of random points from the density (kernel) $nn(\theta)$.
3. Perform importance sampling or the (independence chain) Metropolis-Hastings algorithm using this sample in order to obtain estimates of the characteristics of $p(\theta)$.

Consider a 4-layer feed-forward neural network with functional form:

$$nn(\theta) = eG_2\left(CG_1(A\theta + b) + d\right) + f, \qquad \theta \in \mathbb{R}^n, \tag{16}$$

where A is $H_1 \times n$, b is $H_1 \times 1$, C is $H_2 \times H_1$, d is $H_2 \times 1$, e is $1 \times H_2$ and $f \in \mathbb{R}$. The integers H_1 and H_2 are interpreted as the numbers of cells in the first and second hidden layer of the neural network, respectively. The vector functions $G_1 : \mathbb{R}^{H_1} \to \mathbb{R}^{H_1}$ and $G_2 : \mathbb{R}^{H_2} \to \mathbb{R}^{H_2}$ are defined

by

$$G_1(v) = (g_1(v_1), \cdots, g_1(v_{H_1}))' \quad \text{and} \quad G_2(z) = (g_2(z_1), \cdots, g_2(z_{H_2}))', \quad v \in \mathbb{R}^{H_1}, \;\; z \in \mathbb{R}^{H_2} \tag{17}$$

where $g_1 : \mathbb{R} \to \mathbb{R}$ and $g_2 : \mathbb{R} \to \mathbb{R}$ are the activation functions.

A neural network is used because of its well-known universal approximation property, see e.g. Gallant and White (1989) and Hornik et al. (1989). Stinchcombe (1988) poses a sufficient condition for universal approximation capabilities for hidden layer activation functions other than sigmoid; for example, this condition is satisfied by continuous probability densities. In the following sections, three specifications of (16) will be used:

Type 1 neural network: A standard three-layer feed-forward neural network (in the notation of (16): $H_2 = 1$, $e = 1$, $f = 0$ and g_2 is the identity $g_2(x) = x$, $x \in \mathbb{R}$). As activation function g_1 in (17), we take the scaled arctangent function:

$$g_1(x) = \frac{1}{\pi} \arctan(x) + \frac{1}{2}, \qquad x \in \mathbb{R}. \tag{18}$$

The reason for choosing the arctangent function is that it can be analytically integrated infinitely many times; the scaling is merely done because it is common practice to use activation functions that take values in the unit interval. We show in subsection 3.2.1, that this property makes the neural network, in the role of a density kernel, easy to sample from.

Type 2 neural network: A simplified four-layer network of which the second hidden layer consists of only one cell ($H_2 = 1$, $e = 1$, $f = 0$) and with g_2 the exponential function:

$$g_2(x) = \exp(x), \qquad x \in \mathbb{R}. \tag{19}$$

In this case, the activation g_1 in (17) is taken to be a piecewise-linear function *plin*, defined as:

$$plin(x) = \begin{cases} 0 & x < -1/2 \\ x + 1/2 & -1/2 \le x \le 1/2 \\ 1 & x > 1/2 \end{cases}, \qquad x \in \mathbb{R}. \tag{20}$$

With these activation functions, the neural network function can be analytically integrated (once). We show in subsection 3.2.2, that this property makes Gibbs sampling, see Geman and Geman (1984), possible. To allow for easy sampling it is sufficient to specify a function g_2 which is positive valued and has an analytical expression for its primitive that is analytically invertible; see subsection 3.2.2. Another example of such a function is the logistic function.

Type 3 neural network: A mixture of Student t distributions:

$$nn(\theta) = \sum_{h=1}^{H} p_h \; t(\theta | \mu_h, \Sigma_h, \nu), \tag{21}$$

where p_h ($h = 1, \ldots, H$) are the probabilities of the Student t components and where $t(\theta|\mu_h, \Sigma_h, \nu)$ is a multivariate t density with mode vector μ_h, scaling matrix Σ_h, and ν degrees of freedom:

$$t(\theta|\mu_h, \Sigma_h, \nu) = \frac{\Gamma((\nu+n)/2)}{\Gamma(\nu/2)(\pi\nu)^{n/2}} \, |\Sigma_h|^{-1/2} \left(1 + \frac{(\theta - \mu_h)' \Sigma_h^{-1} (\theta - \mu_h)}{\nu}\right)^{-(\nu+n)/2}. \tag{22}$$

Table 3: Motivation of the particular neural network specifications

specification of $nn(\theta)$	special properties of $nn(\theta)$		consequences of special properties of $nn(\theta)$
Type 1 (3-layer)	- Activation g_1 is analytically integrable infinitely many times.	$\Rightarrow$	- Direct sampling from $nn(\theta)$ is possible.
Type 2 (4-layer)	- Activation g_1 is piecewise-linear. - Activation g_2 is positive valued and analytically integrable, and its primitive is analytically invertible.	$\Rightarrow$	- Gibbs sampling from $nn(\theta)$ is possible.
	- Activation g_2 is the exponential function.	$\Rightarrow$	- Auxiliary variable Gibbs sampling from $nn(\theta)$ is possible.
Type 3 (4-layer)	- $nn(\theta)$ is a mixture of multivariate t densities.	$\Rightarrow$	- Direct sampling from $nn(\theta)$ is possible.

Note that this mixture of t densities is a four-layer feed-forward neural network (with parameter restrictions) in which we have, in the notation of (16), $H_2 = H$ (the number of t densities), $H_1 = Hn$, activation functions

$$g_1(x) = x^2 \qquad \text{and} \qquad g_2(x) = x^{-(\nu+n)} \frac{\Gamma((\nu+n)/2)}{\Gamma(\nu/2)(\pi\nu)^{n/2}}, \qquad x \in \mathbb{R},$$

and weights $e_h = p_h \, |\Sigma_h|^{-1/2}$ $(h = 1, \ldots, H)$, $f = 0$ and:

$$A = \begin{pmatrix} \Sigma_1^{-1/2} \\ \vdots \\ \Sigma_H^{-1/2} \end{pmatrix}, \quad b = \begin{pmatrix} -\Sigma_1^{-1/2}\mu_1 \\ \vdots \\ -\Sigma_H^{-1/2}\mu_H \end{pmatrix}, \quad C = \begin{pmatrix} \iota_n'/\nu & 0 & \cdots & 0 \\ 0 & \iota_n'/\nu & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \iota_n'/\nu \end{pmatrix}, \quad d = \iota_H,$$

where ι_k denotes a $k \times 1$ vector of ones. Notice that $(\theta - \mu_h)'\Sigma_h^{-1}(\theta - \mu_h)$ is the sum of the squared elements of $\Sigma_h^{-1/2}(\theta - \mu_h)$.

The reason for this choice is that a mixture of t distributions is easy to sample from, and that the Student t distribution has fatter tails than the normal distribution.

Table 3 gives an overview of the reasons for which we have chosen these particular specifications. The implications shown in this table will be clarified in the sequel of this paper. Throughout this paper we use the term 'neural network' to denote the classes of functions described above; it should be mentioned here that in part of the literature, see *e.g.* Hastie, Tibshirani and Friedman (2001), such methods are also denoted by 'adaptive basis function methods' or 'dictionary methods', in which a search mechanism is used in order to construct a linear combination of (nonlinear) basis functions that are chosen from a (possibly infinite) set or 'dictionary' of candidate basis functions.

In the next subsections we discuss the three steps of our approach: construction of a neural network, sampling from it, and using the sample in IS or MH.

3.1 Constructing a neural network approximation to a density

First, we discuss a procedure to obtain a Type 1 or Type 2 neural network approximation. Second, we describe a method to construct a Type 3 neural network.

3.1.1 Constructing a Type 1 (3-layer) or Type 2 (4-layer) neural network approximation

We suggest the following procedure to obtain a Type 1 or Type 2 neural network approximation to a certain target density kernel $p(\theta)$. First we draw a set of random uniform points θ^i $(i = 1, \ldots, N)$ in the bounded region to which we restrict the random variable $\theta \in \mathbb{R}^n$ to take its values. Then we approximate the target density kernel $p(\theta)$ with a neural network by minimizing the sum of squared residuals:

$$SSR(A, b, c, d) = \sum_{i=1}^{N} \left(p(\theta^i) - nn\left(\theta^i \middle| A, b, c, d\right)\right)^2, \tag{23}$$

where we use the notation c instead of C, as in our Type 1 and 2 networks this is a $(1 \times H_1)$ vector. We choose the smallest neural network, i.e. the one with the least hidden cells, that still gives a 'good' approximation to the target distribution. One could define a 'good' approximation as one with a high enough squared correlation R^2 between p and nn at the points θ^i $(i = 1, \ldots, N)$.

After that, we check the squared correlation R^2 between the neural network and the target density kernel for a (much) larger set of points than the 'estimation set'. If this R^2 is also high enough, then we say that the approximation is accurate and the estimation set is large enough. In that case the network does not only provide a good approximation to the target density in the points θ^i $(i = 1, \ldots, N)$ but also in between. Otherwise, we increase the number of points N and start all over again. For example, we make the set twice as large. This process continues until the set is large enough to allow the neural network to 'feel' the shape of the target density accurately.

In the case of our Type 1 (three-layer) neural network, we also have to deal with the problem that the neural network function is not automatically non-negative for each θ. In order to try to prevent this we add a penalty term to (23). It should be mentioned that, since a neural network can have a surface that looks like a bed of nails, one should be very careful when checking the non-negativity. For example, one can look for the (global) minimum of $nn(\theta)$ by running a minimization procedure starting with several initial values. Notice that if the minimum of $nn(\theta)$ is a small negative value, one can subtract this negative value from the network's constant d, so that $nn(\theta)$ becomes non-negative for each θ.

In our Type 2 (simplified four-layer) neural network the exponential function, or any positive valued function g_2, implies that non-negativity is automatically taken care of.

3.1.2 Constructing a Type 3 (mixture of t) neural network approximation

We suggest the following procedure to obtain a Type 3 neural network approximation – an adaptive mixture of t densities (AdMit) – to a certain target density kernel $p(\theta)$.

First we compute the mode μ_1 and scale Σ_1 of the first Student t distribution in our mixture as the mode of the target distribution $\mu_1 = \text{argmax}\, p(\theta)$ and Σ_1 as the negative inverse Hessian

of $\log p(\theta)$ evaluated at its mode μ_1. Then we draw a set of points θ^i $(i = 1, \ldots, N)$ from the 'first stage neural network' $nn(\theta) = t(\theta|\mu_1, \Sigma_1, \nu)$, with small ν to allow for fat tails.[6] After that we iteratively add components to the mixture by performing the following steps:

Step 1: Compute the importance sampling weights $w(\theta^i)$ and scaled weights $\tilde{w}(\theta^i)$:

$$w(\theta^i) = \frac{p(\theta^i)}{nn(\theta^i)} \quad \text{and} \quad \tilde{w}(\theta^i) = \frac{w(\theta^i)}{\sum_{i=1}^{N} w(\theta^i)} \qquad (i = 1, \ldots, N).$$

In order to determine the number of components H of the mixture we make use of a simple diagnostic criterium: the coefficient of variation, the standard deviation divided by the mean, of the IS weights $w(\theta^i)$ $(i = 1, \ldots, N)$. If the relative decrease in the coefficient of variation of the importance sampling weights caused by adding one new Student-t component to the candidate mixture is small, e.g. less than 10%, then we stop: the current $nn(\theta)$ is our Type 3 neural network approximation.[7] Otherwise, go to step 2.

Step 2: Add another t distribution with density $t(\theta|\mu_h, \Sigma_h, \nu)$ to the mixture with $\mu_h = \text{argmax } w(\theta) = \text{argmax}\{p(\theta)/nn(\theta)\}$ and Σ_h the negative inverse Hessian of $\log w(\theta) = \log p(\theta) - \log nn(\theta)$ evaluated at its mode μ_h. Here $nn(\theta)$ denotes the latest mixture of $(h-1)$ Student-t densities obtained in the previous iteration of the procedure. An obvious initial value for the maximization procedure for computing $\mu_h = \text{argmax } w(\theta)$ is the point θ^i with the highest weight $w(\theta^i)$ in the sample $\{\theta^i | i = 1, \ldots, N\}$. The idea behind this choice of μ_h and Σ_h is that the new t component should cover a region where the weights $w(\theta)$ are relatively large: the point where the weight function $w(\theta)$ attains its maximum is an obvious choice for the mode μ_h, while the scale Σ_h, the negative inverse Hessian of $\log w(\theta)$ evaluated at its mode μ_h, is the covariance matrix of the local normal approximation to the distribution with density kernel $w(\theta)$ around the point μ_h.

If the region of integration of the parameters θ is bounded, it may occur that $w(\theta)$ attains its maximum at the boundary of the integration region; in this case the negative inverse Hessian of $\log w(\theta)$ evaluated at its mode μ_h may be a very poor scale matrix; in fact this matrix may not even be positive definite. In that case μ_h and Σ_h are obtained as estimates of the mean and covariance matrix of a certain 'residual distribution' with density kernel:

$$res(\theta) = \max\{p(\theta) - c\, nn(\theta), 0\}, \tag{24}$$

where c is a constant; we take $\max\{., 0\}$ to make it a (non-negative) density kernel. These estimates of the mean and covariance matrix of the 'residual distribution' are easily obtained by importance sampling with the current $nn(\theta)$ as the candidate density, using the sample θ^i $(i = 1, \ldots, N)$ from $nn(\theta)$ that we already have. The weights $w_{res}(\theta^i)$ and scaled weights

[6] Throughout this paper we use Student t distributions with $\nu = 1$. There are two reasons for this. First, it enables the methods to deal with fat-tailed target (posterior) distributions. Second, it makes it easier for the iterative procedure by which the Type 3 neural network approximation is constructed to detect modes that are far apart. One could also choose to optimize the degree of freedom of the Student t distributions and/or allow for different degrees of freedom in different Student t distributions. This is a topic for further research.

[7] Notice that $nn(\theta)$ is a proper density, whereas $p(\theta)$ is merely a density kernel. So, the Type 3 neural network does not provide an approximation to the target density kernel $p(\theta)$ in the sense that $nn(\theta) \approx p(\theta)$, but $nn(\theta)$ provides an approximation to the density of which $p(\theta)$ is a kernel in the sense that the ratio $nn(\theta)/p(\theta)$ has relatively little variation.

$\tilde{w}_{res}(\theta^i)$ $(i = 1, \ldots, N)$ are:

$$w_{res}(\theta^i) = \frac{res(\theta^i)}{nn(\theta^i)} = \max\{w(\theta^i) - c, 0\} \quad \text{and} \quad \tilde{w}_{res}(\theta^i) = \frac{w_{res}(\theta^i)}{\sum_{i=1}^{N} w_{res}(\theta^i)}, \tag{25}$$

and μ_h and Σ_h are obtained as:

$$\mu_h = \sum_{i=1}^{N} \tilde{w}_{res}(\theta^i)\theta^i \qquad \Sigma_h = \sum_{i=1}^{N} \tilde{w}_{res}(\theta^i)(\theta^i - \mu_h)(\theta^i - \mu_h)'.$$

There are two issues relevant for the choice of c in (24) and (25). First, the new t density should appear exactly at places where $nn(\theta)$ is too small (relative to $p(\theta)$), i.e. the scale should not be too large. Second, there should be enough points θ^i with $w(\theta^i) > c$ in order to make Σ_h nonsingular. A procedure is to calculate Σ_h for c equal to 100 times the average value of $w(\theta^i)$ $(i = 1, \ldots, N)$; if Σ_h is nonsingular, accept c; otherwise lower c.

Step 3: We now choose the probabilities p_h $(h = 1, \ldots, H)$ in the mixture

$$nn(\theta) = \sum_{h=1}^{H} p_h \; t(\theta|\mu_h, \Sigma_h, \nu),$$

by minimizing the (squared) coefficient of variation of the importance sampling weights. First we draw N points θ_h^i from each component $t(\theta|\mu_h, \Sigma_h, \nu)$ $(h = 1, \ldots, H)$. Then we minimize $E[w(\theta)^2]/E[w(\theta)]^2$, where:

$$E[w(\theta)^k] = \frac{1}{N} \sum_{i=1}^{N} \sum_{h=1}^{H} p_h \; w\big(\theta_h^i\big)^k \qquad (k = 1, 2)$$

with

$$w\big(\theta_h^i\big) = \frac{p(\theta_h^i)}{\sum_{h=1}^{H} p_h \; t\left(\theta_h^i|\mu_h, \Sigma_h, \nu\right)}.$$

Step 4: Draw a sample of N points θ^i $(i = 1, \ldots, N)$ from our new mixture of t distributions:

$$nn(\theta) = \sum_{h=1}^{H} p_h \; t(\theta|\mu_h, \Sigma_h, \nu) \tag{26}$$

and go to step 1; in order to draw a point from (26) we first use a drawing from the $U(0, 1)$ distribution to determine which component $t(\theta|\mu_h, \Sigma_h, \nu)$ is chosen, and then draw from this multivariate t distribution.

It may occur that one is dissatisfied with diagnostics like the weight of the 5% most influential points or the coefficient of variation of the IS weights corresponding to the final candidate density resulting from the procedure above. In that case one may start all over again with a larger number of points N. The idea behind this is that the larger N is, the easier it is for the method to 'feel' the shape of the target density kernel, and to specify the t distributions of the mixture adequately.

Note that an advantage of the Type 3 network, as compared to the Type 1 and 2 networks, is that its construction does not beforehand require the specification of a certain bounded region where the random variable $\theta \in \mathbb{R}^n$ takes its values.

3.2 Sampling from a neural network density

In the following subsections we discuss sampling from Type 1 and Type 2 networks. In the previous subsection we already remarked that sampling from a Type 3 network, a mixture of t densities, only requires a draw from the $U(0,1)$ distribution to determine which component is chosen, and a draw from the chosen multivariate t distribution.

3.2.1 Sampling from a Type 1 (3-layer) neural network density

Suppose the joint density kernel of a certain $\theta \in \mathbb{R}^n$ is given by a standard three-layer feed-forward neural network function with an activation function that is analytically integrable infinitely many times. Since the neural network function is a linear combination of these activation functions, the neural network function itself is integrable infinitely many times.

Hence, the marginal and conditional distribution functions can be evaluated analytically, so that sampling a random vector θ from the density kernel $nn(\theta)$ is easily done by drawing $U(0,1)$ variables and numerically inverting the distribution function; it seems that taking a few steps of the bisection method followed by the Newton-Raphson method works well in practice. In our Type 1 network the activation function is given by the scaled arctangent function in (18), which is analytically integrable infinitely many times. The integrals of the arctangent are given by the following theorem.

Theorem 1: *The n-th integral $J_n(x)$ $(n = 1, 2, \ldots)$ of the arctangent function*

$$J_n(x) \equiv \int \cdots \int \arctan(x) dx \cdots dx \qquad x \in \mathbb{R}$$

is given by

$$J_n(x) = p_n(x) \arctan(x) + q_n(x) \ln(1 + x^2) + r_n(x), \qquad x \in \mathbb{R}, \tag{27}$$

where p_n and q_n are polynomials of degree n and $n-1$, respectively:

$$\begin{aligned} p_n(x) &= p_{n,0} + p_{n,1}\, x + \cdots + p_{n,n-1}\, x^{n-1} + p_{n,n}\, x^n \\ q_n(x) &= q_{n,0} + q_{n,1}\, x + \cdots + q_{n,n-1}\, x^{n-1} \end{aligned}$$

The coefficients $p_{n,k}$ $(k = 0, 1, \ldots, n)$ and $q_{n,k}$ $(k = 0, 1, \ldots, n-1)$ are given by:

$$p_{n,k} = \begin{cases} \frac{(-1)^{(n-k)/2}}{(n-k)!k!} & \text{if } n-k \text{ is even} \\ 0 & \text{if } n-k \text{ is odd} \end{cases} \qquad q_{n,k} = \begin{cases} \frac{(-1)^{(n-k+1)/2}}{2(n-k)!k!} & \text{if } n-k \text{ is odd} \\ 0 & \text{if } n-k \text{ is even} \end{cases} \tag{28}$$

The polynomial r_n (of degree at most $n-1$) plays the role of the integrating constant.

Proof: By induction; see Hoogerheide, Kaashoek and Van Dijk (2004).[8]

This implies that the cumulative distribution function of $\theta \sim nn(\theta)$ where nn is our Type 1 neural network function and where each element θ_i is restricted to a certain interval $[\underline{\theta}_i, \bar{\theta}_i]$

[8] For a particular value of n the validity of Theorem 1 can also be verified by the online Mathematica integration program of Wolfram Research, Inc. on `http://integrals.wolfram.com`

$(i = 1, \ldots, n)$, is given by:

$$CDF_\theta(\tilde{\theta}_1, \ldots, \tilde{\theta}_n) = \left(\frac{1}{2}\sum_{h=1}^{H} c_h + d\right)(\tilde{\theta}_1 - \underline{\theta}_1)\cdots(\tilde{\theta}_n - \underline{\theta}_n)$$
$$+\sum_{h=1}^{H} \frac{c_h}{\pi a_{h1} a_{h2} \cdots a_{hn}} \sum_{D_1=0}^{1}\cdots\sum_{D_n=0}^{1} (-1)^{D_1+\cdots+D_n} J_n\left(\sum_{i=1}^{n} a_{hi}\theta_{i,D_i} + b_h\right). \qquad (29)$$

where we define $\theta_{i,0} = \tilde{\theta}_i$ and $\theta_{i,1} = \underline{\theta}_i$ $(i = 1, 2, \ldots, n)$, the upper and lower bounds of the integration intervals; the primitive $J_n(\cdot)$ is given by (27) in Theorem 1.

The marginal distribution functions $CDF_{\theta_j}(\theta_j)$ $(j = 1, \ldots, n)$ are now obtained by taking $\tilde{\theta}_i = \bar{\theta}_i$ $\forall i = 1, \ldots, n; i \neq j$ in (29). The conditional CDF of θ_j given $\theta_{j+1}, \ldots, \theta_n$ is simply derived by substituting $\sum_{i=j+1}^{n} a_{hi}\theta_i + b_h$ for b_h and treating the neural network as a function of the j-dimensional vector $(\theta_1, \ldots, \theta_j)'$.

3.2.2 Sampling from a Type 2 (4-layer) neural network density

Suppose the joint density kernel of a certain $\theta \in \mathbb{R}^n$ is given by the Type 2 neural network with g_2 the exponential function and g_1 the piecewise-linear function *plin* in (20). It is fairly easy to perform Gibbs sampling from this distribution, as one can divide the (bounded) domain of each θ_i $(i = 1, \ldots, n)$ into a finite number of intervals on which the conditional neural network density is just the exponent of a linear function; the obvious reason for this is that a linear combination of piecewise-linear functions of θ_i is itself a piecewise-linear function of θ_i. Therefore we can analytically integrate the conditional neural network density, and draw from it by analytically inverting the conditional CDF. Note that the three properties of g_2 mentioned below formula (20) are used here explicitly. A more detailed description of this procedure can be found in Hoogerheide, Kaashoek and Van Dijk (2004).

It is also possible to use a different method to draw from a four-layer neural network density: auxiliary variable Gibbs sampling. Using this method, we do not have to restrict ourselves to the piecewise-linear function *plin* when specifying the activation function g_1. It allows for well-known activation functions such as the logistic and scaled arctangent functions. Auxiliary variable Gibbs sampling is a Gibbs sampling technique, developed by Damien et al. (1999). The method is based on work of Edwards and Sokal (1988). In this method, a vector of latent variables u is introduced in an artificial way in order to facilitate drawing from the full set of conditional distributions of θ.

In the case of our Type 2 neural network the vector of latent variables u is $(H \times 1)$ where conditionally on θ the u_h $(h = 1, \ldots, H)$ are independently drawn from uniform distributions:

$$u_h|\theta \;\sim\; U\left(0, \exp\left[c_h\, plin\left(\sum_{i=1}^{n} a_{hi}\theta_i + b_h\right)\right]\right), \qquad h = 1, \ldots, H. \qquad (30)$$

The elements θ_i $(i = 1, \ldots, n)$ are drawn conditionally on u and θ_{-i}, the set of all other elements of θ, from the uniform distribution on the interval $[\theta_{i,LB}(u, \theta_{-i}), \theta_{i,UB}(u, \theta_{-i})]$, where:

$$\theta_{i,LB}(u, \theta_{-i}) \;=\; \max\left\{\underline{\theta}_i, \max_{1\leq h\leq H}\left\{\frac{1}{a_{hi}}\left(\frac{\log(u_h)}{c_h} - \frac{1}{2} - \left(\sum_{j=1, j\neq i}^{n} a_{hj}\theta_j + b_h\right)\right)\right| \right.$$
$$\left. c_h a_{hi} > 0, 0 < \frac{\log(u_h)}{c_h} < 1\right\}\right\}, \qquad (31)$$

$$\theta_{i,UB}(u,\theta_{-i}) \;=\; \min\left\{\bar{\theta}_i,\ \min_{1\le h\le H}\left\{\frac{1}{a_{hi}}\left(\frac{\log(u_h)}{c_h}-\frac{1}{2}-\left(\sum_{j=1,j\neq i}^{n} a_{hj}\theta_j+b_h\right)\right)\right|\right.$$
$$\left. c_h a_{hi}<0, 0<\frac{\log(u_h)}{c_h}<1\right\}\right\}, \quad (32)$$

where $[\underline{\theta}_i, \bar{\theta}_i]$ is the interval to which θ_i $(i = 1, \ldots, n)$ is a priori restricted. The derivations of these conditional distributions are given in Hoogerheide, Kaashoek and Van Dijk (2004).

3.3 Importance sampling and Metropolis-Hastings

Once we have obtained a sample of random drawings from the neural network density $nn(\theta)$, we use this sample in order to estimate those characteristics of the target density $p(\theta)$ that we are interested in. Two methods that we can use for this purpose are importance sampling and the Metropolis-Hastings algorithm. A discussion of importance sampling can be found in Bauwens et al. (1999). The Metropolis-Hastings (MH) algorithm was introduced by Metropolis et al. (1953) and generalized by Hastings (1970).
Note that in the case of a four-layer neural network we need Gibbs sampling in order to obtain the sample, so that the consecutive drawings are not independent. This case can be dealt with using a Metropolis-Hastings within Gibbs algorithm, in which a MH step is considered after each time an element θ_i is drawn from its conditional neural network distribution. So, we have the following eight 'neural network based' algorithms at hand:

- Neural Network Importance Sampling (NNIS) and Neural Network Metropolis-Hastings (NNMH) in which IS or MH is performed using random vectors that are (directly) drawn from a 3-layer neural network;
- Gibbs Neural Network Importance Sampling (GiNNIS) and Gibbs with Auxiliary Variables Neural Network Importance Sampling (GiAuVaNNIS) in which IS is performed using random vectors that are drawn from a 4-layer neural network by Gibbs sampling (possibly with auxiliary variables);
- Gibbs Neural Network Metropolis-Hastings (GiNNMH) and Gibbs with Auxiliary Variables Neural Network Metropolis-Hastings (GiAuVaNNMH) in which Metropolis-Hastings within Gibbs is performed using random vectors that are drawn from a 4-layer neural network by Gibbs sampling (possibly with auxiliary variables);
- IS or MH using random vectors that are (directly) drawn from an Adaptive Mixture of t distributions (AdMit-IS or AdMit-MH).

Table 4 gives an overview.

4 Example I: Neural Network sampling methods applied to a bivariate conditionally normal distribution

In this section we consider an illustrative bivariate distribution in order to show the feasibility of the neural network approach and to compare the performance of the different neural network based methods. In the notation of the previous section we have $\theta = (X_1, X_2)'$.

Table 4: Overview of neural network based sampling algorithms

	Importance sampling	Metropolis-Hastings
Type 1 (3-layer) neural network: direct sampling	NNIS	NNMH
Type 2 (4-layer) neural network: (auxiliary variable) Gibbs sampling	Gi(AuVa)NNIS	Gi(AuVa)NNMH
Type 3 neural network (adaptive mixture of t densities): direct sampling	AdMit-IS	AdMit-MH

Let X_1 and X_2 be two random variables, for which X_1 is normally distributed given X_2 and vice versa. Then the joint distribution, after location and scale transformations in each variable, can be written as (see Gelman and Meng (1991)):

$$p(x_1, x_2) \propto \exp\left(-\frac{1}{2}\left[Ax_1^2x_2^2 + x_1^2 + x_2^2 - 2Bx_1x_2 - 2C_1x_1 - 2C_2x_2\right]\right), \tag{33}$$

where A, B, C_1 and C_2 are constants. We consider the symmetric case in which $A = 1$, $B = 0$, $C_1 = C_2 = 3$, with conditional distributions

$$X_1|X_2 = x_2 \sim N\left(\frac{3}{1+x_2^2}, \frac{1}{1+x_2^2}\right) \qquad X_2|X_1 = x_1 \sim N\left(\frac{3}{1+x_1^2}, \frac{1}{1+x_1^2}\right). \tag{34}$$

For the Type 1 and 2 networks, we restrict the variables X_1 and X_2 to the interval [-2.5,7.5], i.e. we only consider the region

$$\{(X_1, X_2)| -2.5 \leq X_1 \leq 7.5, -2.5 \leq X_2 \leq 7.5\}. \tag{35}$$

This restriction does not affect our estimates, as the probability mass outside this region is negligible.

The contourplots of the neural network approximations[9] are given by Figure 7, together with the contourplot of the target density. These contourplots confirm that the three classes of neural networks are able to provide reasonable approximations to the target density. Figure 8 illustrates how the AdMit procedure iteratively constructs an approximating (mixture of t) candidate density in four steps.

After we have constructed neural network approximations, we sample from these networks and use the samples in IS or MH. Many diagnostic checks have been developed for assessing the

[9]We constructed a Type 1 network with $H = 50$, $R^2 = 0.9966$ on its training set of 1000 points, and $R^2 = 0.9936$ on its test set of 5000 points. We obtained a Type 2 network with $H = 13$, $R^2 = 0.9944$ on its training set of 1000 points, and $R^2 = 0.9756$ on its test set of 5000 points. We also constructed a mixture of four t distributions with a sample of 1000 IS weights with coefficient of variation equal to 0.840 (and in which the 5% most influential points have 11.6% weight).

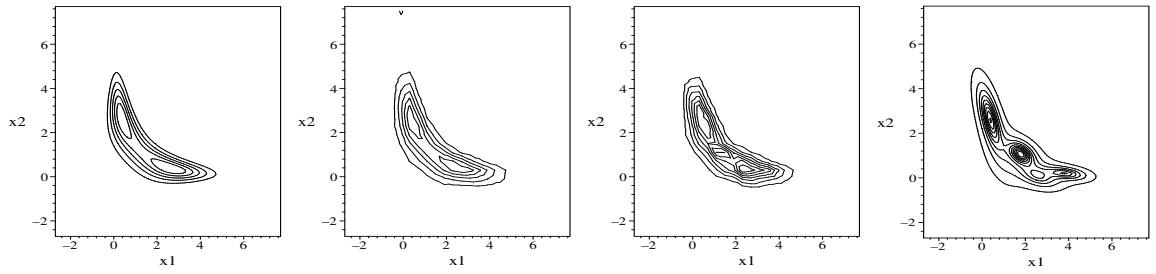

Figure 7: Contour plots: conditionally normal bivariate distribution in (34) (left), and its Type 1 (second), Type 2 (third), and Type 3 (right) neural network approximation

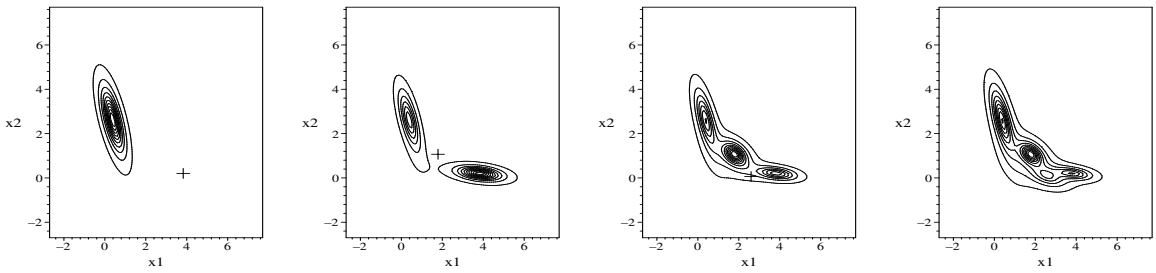

Figure 8: Illustration of the AdMit procedure for constructing a Type 3 (mixture of t) neural network approximation to a target (posterior) density: in four steps a candidate density is constructed; the cross denotes the point at which the weight function $p(x_1, x_2)/nn(x_1, x_2)$ corresponding to the displayed candidate density $nn(x_1, x_2)$ attains its maximum. For the four shown candidate densities the coefficient of variation of the importance sampling weights is 4.01, 1.39, 0.93, 0.87, respectively.

convergence of the IS or MH method; see *e.g.* Geweke (1989) for the IS method and Cowles and Carlin (1996) and Brooks and Roberts (1998) for MCMC methods. Here we use the following simple heuristic rule to obtain estimates of the means with a precision of 1 decimal: for each algorithm we construct two samples, and we say that convergence has been achieved if the difference between the two estimates of $E(X_1)$ and the difference between the two estimates of $E(X_2)$ are both less than 0.05.[10] The results are in Table 5. Note that the eight neural network sampling algorithms all yield estimates of $E(X_1)$ and $E(X_2)$ differing less than 0.05 from the real values. The table shows numerical standard errors and the corresponding relative numerical efficiency (RNE), see Geweke (1989). The numerical standard errors are estimates of the standard deviations of the IS estimators of $E(X_1)$ and $E(X_2)$. The RNE is the ratio between the IS estimator's estimated variance and (an estimate of) the variance that an estimator based on direct sampling would have (with the same number of drawings). The RNE is an indicator of the efficiency of the chosen importance function; in the ideal case where target and importance density coincide the RNE equals one, whereas a very poor importance density will have an RNE close to zero.

The total weight of the 5% most influential points is below 15% for the three IS algorithms

[10]The number of drawings required may depend on an initial value such as the seed of the random number generator; for each algorithm the experiment has been repeated several times and the results are robust in the sense that in most cases convergence had been reached after the reported number of drawings.

and the values of the RNE are rather high, confirming the quality of the importance density. The rather high MH acceptance rates above 50% indicate the quality of the neural network as a candidate density in MH.

If we look at the computing times (on an AMD Athlon^{TM} 1.4 GHz processor) required for generating the samples, we conclude that AdMit-IS and AdMit-MH are the winners in this example. In AdMit-IS or AdMit-MH the construction of the network, the sampling, and the IS or MH require altogether 2.0 seconds, whereas the other methods take much more time to construct a network and to generate an adequate sample.

The NNIS and NNMH algorithms are relatively slow, as relatively many hidden cells ($H = 50$) are required to provide a reasonable Type 1 neural network approximation, which makes optimization rather time consuming; also sampling from a Type 1 network is rather slow as this requires a numerical method, such as the Newton method, in order to perform the inverse transformation method. Quicker optimization methods for the Type 1 and 2 neural networks are a topic for further research.

The GiAuVaNNIS and GiAuVaNNMH algorithms are slightly slower than the GiNNIS and GiNNMH methods; although drawing a point takes more time in the latter methods, the introduction of the auxiliary variables increases the serial correlation in the Gibbs sequence in such a way that many more drawings are required to reach convergence.

Table 5: Neural network based sampling results for the conditionally normal bivariate distribution in (34)

	real values	NNIS	NNMH	GiNNIS	GiNNMH	GiAuVa NNIS	GiAuVa NNMH	AdMit IS	AdMit MH
$E(X_1)$	1.459	1.487	1.504	1.472	1.433	1.468	1.477	1.464	1.467
(num. std. error)		(0.019)						(0.015)	
[RNE]		[0.896]						[0.649]	
$E(X_2)$	1.459	1.450	1.434	1.444	1.490	1.454	1.436	1.459	1.458
(num. std. error)		(0.019)						(0.016)	
[RNE]		[0.885]						[0.619]	
$\sigma(X_1)$	1.234	1.239	1.247	1.233	1.229	1.239	1.237	1.236	1.245
$\sigma(X_2)$	1.234	1.239	1.235	1.223	1.244	1.233	1.234	1.242	1.235
$\rho(X_1, X_2)$	-0.760	-0.764	-0.766	-0.755	-0.757	-0.758	-0.757	-0.759	-0.759
total time		257.0 s	257.0 s	66.5 s	79.9 s	81.3 s	85.6 s	2.0 s	2.0 s
time construction NN		225.2 s	225.2 s	62.6 s	62.6 s	62.6 s	62.6 s	1.1 s	1.1 s
time sampling		31.8 s	31.8 s	3.9 s	17.3 s	18.7 s	23.0 s	0.9 s	0.9 s
drawings		5000	5000	10000	40000	80000	80000	10000	10000
time/draw		6.4 ms	6.4 ms	0.39 ms	0.43 ms	0.23 ms	0.29 ms	0.09 ms	0.09 ms
5% weights		6.3 %		7.2 %		7.2 %		12.9 %	
coeff. var. weights		0.382		0.239		0.251		0.840	
acc. rate			84.6%		90.0 %		92.7 %		52.7 %
serial corr. X_1			0.15	0.65	0.73	0.90	0.92		0.45
serial corr. X_2			0.14	0.67	0.72	0.84	0.86		0.45

5 Example II: Neural Network sampling methods applied to posterior distributions in a simple IV regression model

In this section we consider posterior distributions in IV regression models in order to compare the performance of the Type 3 (mixture of t) neural network sampling method (AdMit) with some other sampling methods.

First, consider the joint posterior of π and β in (9) for the data set simulated from the model (2) - (3) with $\pi = 0.1$ (weak identification) and $\rho = 0.99$ (strong endogeneity) truncated to the region

$$\{(\pi, \beta)| -0.25 \leq \pi \leq 0.25, -10 \leq \beta \leq 10\}. \tag{36}$$

Figure 2 shows its contourplot on this region (36).

The contourplot of the Type 3 neural network approximation[11] is given by Figure 9, together with the contourplot of the target density. This contourplot confirms that this class of neural networks is able to provide reasonable approximations to a wide class of (possibly multi-modal) target densities.

In this example the Gibbs sampler failed: the Gibbs sequence remained in one of the two ridges for at least 100 million drawings, yielding a scatter plot like in Figure 9. Of course, one can draw from the other ridge by choosing a different initial value (in or close to the other ridge), but it is not a trivial issue how to weight the results from the two ridges, i.e. one has to determine which part of the posterior probability mass is contained in each of both ridges.

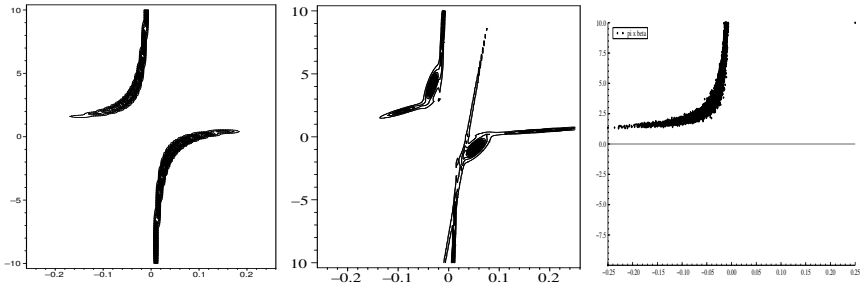

Figure 9: Contourplots in the $\pi \times \beta$ plane: joint posterior of π and β in IV model for simulated data set with $\pi = 0.1$, $\rho = 0.99$ (left), and its Type 3 neural network approximation (middle); scatter plot of sample obtained by Gibbs sampler (right)

Second, we consider the joint posterior of $\pi = (\pi_1, \pi_2)'$ and β in (9) with $k = 2$ instruments for $T = 50$ simulated data from the model (2) - (3) with $\pi_1 = \pi_2 = 0.1$ (weak identification) and $\rho = 0.99$ (strong endogeneity) truncated to the region

$$\{(\pi_1, \pi_2, \beta)| -0.5 \leq \pi_i \leq 0.5 \ (i = 1, 2), -10 \leq \beta \leq 10\}. \tag{37}$$

Figure 10 shows the shape of a credible set on this region (37), together with the shapes of credible sets in similar models with $T = 50$ simulated data from the model (2) - (3) with $\pi_1 = \pi_2 = 0$ (no identification) and $\pi_1 = \pi_2 = 1$ (strong identification). Note that the same shapes that showed up in the 2-dimensional distributions (ridges, bimodality and nearly elliptical shapes) also occur in these 3-dimensional distributions.

We construct a Type 3 neural network approximation, a mixture of 15 Student t distributions, and use 1000000 drawings from it in IS and MH; see Table 6.

We compare its performance (in the same computing time) with IS using a unimodal importance density, the Student t distribution with $\nu = 1$ degree of freedom. In order to give the unimodal density a fair chance, we first take 4 steps in which the mode and scale are updated as the estimated mean and covariance matrix of the target distribution in the previous step. The

[11] We constructed a mixture of 8 Student t distributions with a sample of 50000 IS weights with coefficient of variation of 2.1.

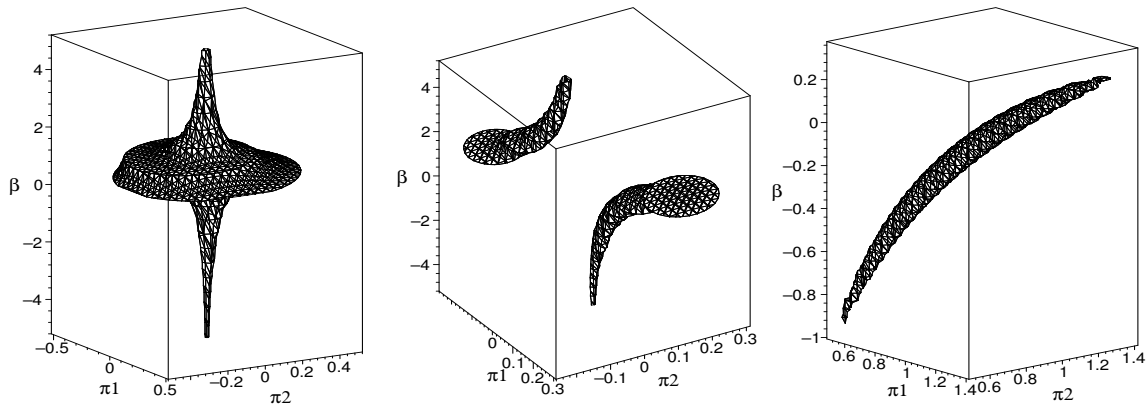

Figure 10: Credible sets for parameters π_1, π_2, β in IV model (2) - (3) for simulated data sets from this model with strong endogeneity ($\rho = 0.99$) and no ($\pi_1 = \pi_2 = 0$), weak ($\pi_1 = \pi_2 = 0.1$) or strong ($\pi_1 = \pi_2 = 1$) identification, respectively.

results are in Table 6. If we compare the numerical standard errors, AdMit-IS gives estimates of $E(\pi_1)$, $E(\pi_2)$ and $E(\beta)$ with standard errors that are 1.9, 1.9 and 3.3 times as small, respectively. Also notice the huge differences between the RNEs (especially for the estimate of $E(\beta)$) in the two IS methods.

We also compare the performance of AdMit-IS with the random walk (RW) Metropolis-Hastings algorithm with candidate steps from a Student t distribution with $\nu = 1$ degree of freedom. Again, we first take 4 steps in which the scale is updated as the estimated covariance matrix of the target distribution in the previous step. The results are in Table 6. We have repeated the RW MH algorithm 5 times: the standard deviations of the estimates of $E(\pi_1)$, $E(\pi_2)$ and $E(\beta)$ are 3.5 10^{-4}, 3.9 10^{-4} and 0.0130, respectively; the AdMit standard errors are 2.2, 2.5 and 2.0 times as small as these standard deviations.

The Gibbs sampler failed in this example: the Gibbs sequence remained in one of the two ridges for 25000000 drawings (taking 1039 s).

We conclude that in this example the AdMit approach outperforms three competing algorithms.

Finally, consider the joint posterior of π and β in (9) for the data set simulated from the model (2) - (3) with $k = 1$ instrument with $\pi = 1$ (strong identification) and $\rho = 0$ (no endogeneity), truncated to the region

$$\{(\pi, \beta)| -0.5 \leq \pi \leq 1.5, -10 \leq \beta \leq 10\}. \tag{38}$$

Figure 2 shows its contourplot, which shows an elliptical shape.

We construct a Type 3 neural network approximation, a mixture of 2 Student t distributions. Again, we use a simple heuristic rule to obtain estimates of the means with a precision of 2 decimals: for each algorithm we construct two samples, and we say that convergence has been achieved if the difference between the two estimates of $E(\pi)$ and the difference between the two estimates of $E(\beta)$ are both less than 0.005.[12] The results are in Table 7.

[12]The number of drawings required may depend on an initial value such as the seed of the random number

Table 6: Sampling results for the non-elliptically shaped posterior distribution in the IV regression (2) - (3) with $k = 2$ instruments for simulated data with $\pi = (0.1, 0.1)'$ (weak identification), $\rho = 0.99$ (strong endogeneity)

	real values	AdMit IS	AdMit MH	adaptive t_1 IS	adaptive RW MH
$E(\pi_1)$	0.0199	0.200	0.195	0.0203	0.0206
(num. std. error)		($1.57\ 10^{-4}$)		($3.0\ 10^{-4}$)	
[RNE]		[0.3622]		[0.0032]	
$E(\pi_2)$	0.0157	0.0158	0.0153	0.0161	0.0165
(num. std. error)		($1.56\ 10^{-4}$)		($2.9\ 10^{-4}$)	
[RNE]		[0.3586]		[0.0034]	
$E(\beta)$	0.6404	0.6357	0.6531	0.6039	0.6121
(num. std. error)		(0.0065)		(0.0215)	
[RNE]		[0.2211]		[0.00067]	
$\sigma(\pi_1)$	0.0946	0.0945	0.0943	0.0948	0.0946
$\sigma(\pi_2)$	0.0935	0.0934	0.0934	0.0936	0.0935
$\sigma(\beta)$	3.0643	3.0745	3.0713	3.0506	3.0816
total time		921 s	921 s	1030 s	1138 s
time construction NN		598 s	598 s		
time adapting scale				83 s	83 s
time sampling		323 s	323 s	947 s	1055 s
drawings		$1\ 10^6$	$1\ 10^6$	$30\ 10^6$	$50\ 10^6$
time/draw		0.32 ms	0.32 ms	0.03 ms	0.02 ms
5% weights		27.3 %		99.999 %	
coeff. var. weights		1.47		21.6	
acc. rate			32.5 %		2.3 %
serial corr. π_1			0.66		0.994
serial corr. π_2			0.66		0.994
serial corr. β			0.72		0.996

We compare AdMit's performance with the Gibbs sampler, the random walk MH algorithm with candidate steps from a t_1 distribution with scale matrix equal to the negative inverse Hessian of the log-posterior kernel evaluated at its mode, and IS/MH with a t_1 or normal candidate density around the mode of the target distribution. In this case of an elliptical (posterior) target distribution the methods using a unimodal candidate density all perform well. Although the neural network approach is feasible in this example, it is slower than several competing algorithms. This stresses that different sampling methods dominate in different cases; the neural network approach is especially useful for target densities with non-elliptical contours. The development of strategies to determine which method should be used in which situation is a topic for further research.

generator; for each algorithm the experiment has been repeated several times and the results are robust in the sense that in most cases convergence had been reached after the reported number of drawings.

Table 7: Sampling results for the elliptically shaped posterior distribution in the IV regression (2) - (3) with $k = 1$ instruments for simulated data with $\pi = 1$ (strong identification) and $\rho = 0$ (no endogeneity)

	real values	AdMit IS	AdMit MH	Gibbs	RW MH	RW MH adaptive	IS t_1	MH t_1	IS normal	MH normal
$E(\pi)$	0.908	0.908	0.911	0.910	0.908	0.907	0.908	0.911	0.909	0.909
(num. std. error)		(0.004)					(0.004)		(0.001)	
(RNE)		(0.691)					(0.691)		(0.910)	
$E(\beta)$	-0.028	-0.025	-0.029	-0.029	-0.029	-0.027	-0.025	-0.032	-0.026	-0.027
(num. std. error)		(0.004)					(0.004)		(0.002)	
(RNE)		(0.668)					(0.668)		(0.863)	
$\sigma(\pi)$	0.089	0.093	0.089	0.091	0.090	0.090	0.093	0.088	0.087	0.087
$\sigma(\beta)$	0.106	0.105	0.102	0.104	0.105	0.106	0.105	0.105	0.102	0.102
$\rho(\pi, \beta)$	0.017	0.041	-0.013	0.086	0.021	0.041	0.041	0.015	-0.019	-0.020
total time		20.8 s	20.9 s	0.03 s	0.64 s	1.28 s	0.03 s	0.11 s	0.11 s	0.12 s
time construction NN		20.7 s	20.7 s							
time adapting scale						0.64 s				
time sampling		0.05 s	0.16 s	0.03 s	0.64 s	0.64 s	0.03 s	0.11 s	0.11 s	0.12 s
drawings		1000	2500	1000	40000	40000	1000	2500	4000	4000
time/draw		0.05 ms	0.06 ms	0.03 ms	0.02 ms	0.02 ms	0.03 ms	0.04 ms	0.03 ms	0.03 ms
5% weights		11.1 %					11.1 %		7.5 %	
coeff. var. weights		0.797					0.797		0.163	
acc. rate			58.6 %		39.0 %	38.0 %		60.5 %		93.5 %
serial corr. π			0.40	-0.02	0.85	0.85		0.38		0.11
serial corr. β			0.39	-0.04	0.85	0.85		0.36		0.14

6 Conclusion

In this paper we have shown that the shape of Bayesian credible sets is often non-elliptical in instrumental variable regression models with weak instruments and/or strong endogeneity. Structural inference is possible but the credible sets may indicate large uncertainty. Unless one uses a truncated region of integration, implied reduced form inference is not possible due to an improper posterior. This has important implications for forecasting and policy analysis.

In order to accurately approximate the shape of such non-elliptical credible sets we have introduced a class of neural network sampling algorithms. In these algorithms neural network functions are used as an importance or candidate density in importance sampling or the Metropolis-Hastings algorithm. Neural networks are natural importance or candidate densities, as they have a universal approximation property and are easy to sample from. We have shown how to sample from three types of neural networks. One can sample directly from a certain 3-layer network. Using a 4-layer network one can, depending on the specification of the network, either use a Gibbs sampling approach or sample directly from a mixture of distributions. A key step in the proposed class of methods is the construction of a neural network that approximates the target density accurately. The methods have been tested on an illustrative example; the 4-layer network specified as the mixture of t distributions performed the best among the proposed sampling procedures. In another experiment concerning a bimodal posterior distribution in an IV regression for a simulated data set the approach using a mixture of t distributions provided (in the same computing time) more accurate results than IS with a unimodal importance density or a random walk Metropolis-Hastings algorithm, whereas the Gibbs sampler failed in this example. These results indicate the feasibility and the possible usefulness of the neural network approach. We emphasize that it is naive to expect one sampling method to dominate in all practical cases. We emphasize that one needs to develop a strategy in which a sophisticated network is specified

for complex, non-elliptical densities, while in a relatively simple case of near-elliptical contours a unimodal density or a bimodal mixture may be sufficiently accurate as a candidate density. Clearly, more work is needed in this area and will be reported in future work.

We end this paper with some remarks on how to apply and to extend the proposed techniques. First, one may use these results in model selection and model averaging and investigate the effect of using accurate non-elliptical credible sets instead of naive or asymptotic sets.

Second, one may consider other ways of specifying and estimating neural networks. We mention here the following possible extensions. One may pursue the construction of well-behaved neural networks with other activation functions which are more smooth than the piecewise-linear one. We noted in section 2 that it is possible to perform auxiliary variable Gibbs sampling from a 4-layer neural network density with a logistic function or scaled arctangent instead of the piecewise-linear function. One may also investigate the effects of substituting the exponential function in the second hidden layer by a different function such as the logistic function. One may also, as a first step, transform the posterior density function to a more regular shape. This line of research is recently pursued by *e.g.* Bauwens, Bos, Van Dijk and Van Oest (2004) in a class of adaptive direction sampling methods using radial-basis functions (ARDS). A combination of ADS and neural network sampling may be of interest. In practice, one encounters cases where only part of the posterior density is ill-behaved. Then one may combine the neural network approach for the 'difficult part' with a Gibbs sampling approach for the regular part of the model. In recent work Richard (1998) and Liesenfeld and Richard (2002) constructed an efficient importance sampling technique where the estimation of the parameters of the importance function is done in a sequence of optimization steps. Another area of further research is to consider different flexible candidate density functions involving Hermite polynomials, see *e.g.* Gallant and Tauchen (1993) and the references cited there. Also, more sophisticated Monte Carlo methods like bridge sampling, see *e.g.* Meng and Wong (1996) and Frühwirth-Schnatter (2004), may be explored in combination with neural networks. We intend to report on this in future work.

Third, more experience is needed with empirical econometric models like the models of local average treatment effects, see Imbens and Angrist (1994) or the business cycle models as specified by Hamilton (1989) and Paap and Van Dijk (2003), or stochastic volatility models as given by Shephard (1996), and dynamic panel data models; see Pesaran and Smith (1995).

Fourth, the neural network approximations proposed in this paper may be useful for modelling such processes as volatility in financial series, see e.g. Donaldson and Kamstra (1997), and for evaluating option prices, see Hutchinson, Lo and Poggio (1994). We intend to report on this in future research.

A Derivation of the conditional and marginal posterior densities of the structural parameter β and the reduced form parameter π in a simple IV regression

In order to derive the conditional posterior density for β from the joint density kernel (8) we apply the following decomposition[13] to the determinant $|(\varepsilon\ v)'(\varepsilon\ v)|$:

$$|(\varepsilon\ v)'(\varepsilon\ v)| = |v'v||\varepsilon' M_v \varepsilon|. \tag{39}$$

It follows that

$$p(\beta, \pi | y_1, y_2, X) \propto |(y_1 - y_2\beta)' M_v (y_1 - y_2\beta)|^{-T/2} |v'v|^{-T/2}. \tag{40}$$

[13]This decomposition is Theorem A.3.2 (p. 594) of Anderson (1984) with $A_{11} = \varepsilon'\varepsilon$, $A_{12} = A_{21}' = \varepsilon' v$,$A_{22} = v'v$.

We rewrite the sum of squares:

$$(y_1 - y_2\beta)'M_v(y_1 - y_2\beta) = (T-1)s^2_{\hat{\beta}} + (\beta - \hat{\beta})'y_2'M_v y_2(\beta - \hat{\beta}) \tag{41}$$

where we define $\hat{\beta} \equiv (y_2'M_v y_2)^{-1}(y_2'M_v y_1)$, and $(T-1)s^2_{\hat{\beta}} \equiv (y_1 - y_2\hat{\beta})'M_v(y_1 - y_2\hat{\beta})$, the sum of squared residuals in a regression of $M_v y_1$ on $M_v y_2$, which are – by the definition of the 'residual maker' M_v and Frisch-Waugh – the residuals in a regression of y_1 on y_2 and v.
It follows from (41) that the joint posterior density kernel in (40) can be written as:

$$p(\beta, \pi|y_1, y_2, X) \propto \left[(T-1)s^2_{\hat{\beta}}\right]^{-T/2} \left[1 + \tfrac{1}{T-1}\tfrac{(\beta-\hat{\beta})^2}{s^2_{\hat{\beta}}(y_2'M_v y_2)^{-1}}\right]^{-T/2} |v'v|^{-T/2} \tag{42}$$

It immediately follows from (42) that the conditional distribution of β given π is the (univariate) Student t distribution with mode $\hat{\beta}$, scale $s^2_{\hat{\beta}}(y_2'M_v y_2)^{-1}$ and $(T-1)$ degrees of freedom with density given by (10).

It follows in an analogous fashion like (39)-(42) that the joint posterior density kernel can be written as:

$$\begin{aligned} p(\beta, \pi|y_1, y_2, X) \propto\ & \left[(T-k)s^2_{\hat{\pi}}\right]^{-T/2} \times \\ & \times \left[1 + \frac{1}{T-k}(\pi - \hat{\pi})'(s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1})^{-1}(\pi - \hat{\pi})\right]^{-T/2} |\varepsilon'\varepsilon|^{-T/2} \end{aligned} \tag{43}$$

where $\hat{\pi} \equiv (X'M_\varepsilon X)^{-1}(X'M_\varepsilon y_2)$ and $(T-k)s^2_{\hat{\pi}} \equiv (y_2 - X\hat{\pi})'M_\varepsilon(y_2 - X\hat{\pi})$, the sum of squared residuals in a regression of $M_\varepsilon y_2$ on $M_\varepsilon X$, which are the residuals in a regression of y_2 on X and ε, so that the conditional distribution of π given β is (k-dimensional) Student t with mode $\hat{\pi}$, scaling $s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1}$ and $(T-k)$ degrees of freedom with density given by (12).

We obtain the marginal posterior density of β by dividing the joint posterior density of (β, π) in (43) by the conditional density of π given β in (12):

$$\begin{aligned} p(\beta|y_1, y_2, X) &= \frac{p(\beta, \pi|y_1, y_2, X)}{p(\pi|\beta, y_1, y_2, X)} \propto \frac{\left[(T-k)s^2_{\hat{\pi}}\right]^{-T/2} |\varepsilon'\varepsilon|^{-T/2}}{|s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1}|^{-1/2}} \\ &= |X'M_\varepsilon X|^{-1/2} \left[(T-k)s^2_{\hat{\pi}}\right]^{(T-k)/2} |\varepsilon'\varepsilon|^{-T/2}, \end{aligned} \tag{44}$$

Recall that $(T-k)s^2_{\hat{\pi}}$ is defined as the sum of squared residuals in a regression of y_2 on X and ε:

$$(T-k)s^2_{\hat{\pi}} = (M_X y_2)'M_{M_X\varepsilon}M_X y_2 = (M_X\varepsilon)'M_{M_X y_2}M_X\varepsilon \ \frac{y_2'M_X y_2}{\varepsilon'M_X\varepsilon} \tag{45}$$

where we have again used a decomposition like (39). The term $(M_X\varepsilon)'M_{M_X y_2}M_X\varepsilon$ in (45) is the sum of squared residuals in a regression of ε on X and y_2, which is equal to

$$(M_{y_2}\varepsilon)'M_{M_{y_2}X}M_{y_2}\varepsilon = (M_{y_2}y_1)'M_{M_{y_2}X}M_{y_2}y_1, \tag{46}$$

as $\varepsilon \equiv y_1 - y_2\beta$ and $M_{y_2}y_2 = 0$. From (45) and (46) we have $(T-k)s^2_{\hat{\pi}} \propto (\varepsilon'M_X\varepsilon)^{-1}$. Substituting

$$(T-k)s^2_{\hat{\pi}} \propto (\varepsilon'M_X\varepsilon)^{-1} \quad \text{and} \quad |X'M_\varepsilon X| \propto \frac{\varepsilon'M_X\varepsilon}{\varepsilon'\varepsilon}, \tag{47}$$

where the latter immediately follows from a decomposition like (39), into (44) yields the marginal posterior density kernel of β in (11), the ratio of two Student t kernels.

We obtain the marginal posterior density of π by dividing the joint posterior density of (β, π) in (42) by the conditional density of β in (10):

$$\begin{aligned} p(\pi|y_1, y_2, X) &= \frac{p(\beta, \pi|y_1, y_2, X)}{p(\beta|\pi, y_1, y_2, X)} \propto \frac{\left[(T-1)s^2_{\hat{\beta}}\right]^{-T/2} |v'v|^{-T/2}}{|s^2_{\hat{\beta}}(y_2' M_v y_2)^{-1}|^{-1/2}} \\ &= |y_2' M_v y_2|^{-1/2} \left[(T-1)s^2_{\hat{\beta}}\right]^{-(T-1)/2} |v'v|^{-T/2}. \end{aligned} \tag{48}$$

In a similar way like the derivation of (47) it can be derived that:

$$(T-1)s^2_{\hat{\beta}} \propto \frac{v' M_{[y_1\ y_2]} v}{v' M_{y_2} v} \quad \text{and} \quad |y_2' M_v y_2| \propto \frac{v' M_{y_2} v}{v'v}. \tag{49}$$

Since $M_{y_2}\ y_2 = M_{[y_1\ y_2]}\ y_2 = 0$ we have $M_{y_2} v = -M_{y_2} X\pi$, $M_{[y_1\ y_2]}\ v = -M_{[y_1\ y_2]}\ X\pi$, so that substituting (49) into (48) yields the marginal posterior density kernel of π in (13)-(14), the ratio of a product of two Student t kernels in the numerator and one Student t kernel in the denominator

References

[1] Anderson, T.W. (1984): *An Introduction to Multivariate Statistical Analysis*, second edition, Wiley, New York.

[2] Angrist, J.D., G.W. Imbens and D.B. Rubin(1996): "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association* 91, 444-455.

[3] Angrist, J.D. and A.B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?", *Quarterly Journal of Economics*, 106, 979-1014.

[4] Bauwens, L. and H.K. van Dijk (1990): "Bayesian limited information analysis revisited". In: J. J. Gabszewicz et al. (eds), *Economic Decision-Making: Games, Econometrics and Optimisation*, North-Holland, Amsterdam.

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

[6] Bauwens, L., C.S. Bos, H.K. van Dijk and R.D. van Oest (2004): "Adaptive Radial-Based Direction Sampling: Some flexible and robust Monte Carlo integration methods", *Journal of Econometrics*.

[7] Bos, C.S., R.J. Mahieu and H.K. van Dijk (2000): "Daily Exchange Rate Behaviour and Hedging of Currency Risk", *Journal of Applied Econometrics* 15, 671-696.

[8] Brooks, S.P. and G.O. Roberts (1998): "Convergence Assessment Techniques for Markov Chain Monte Carlo", *Statistics and Computing* 8, 319-335.

[9] Chib, S. and E. Greenberg (1996): "Markov Chain Monte Carlo Simulation Methods in Econometrics", *Econometric Theory*, 12(3), 409-431.

[10] Cowles, M.K. and B.P. Carlin (1996): "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review", *Journal of the American Statistical Association* 91, 883-904.

[11] Damien, P., J. Wakefield and S. Walker (1999), "Gibbs Sampling for Bayesian Non-conjugate and Hierarchical Models by using Auxiliary Variables", *Journal of the Royal Statistical Society B*, 61, 331-344

[12] Donaldson, R.G. and M. Kamstra (1997), "An Artificial Neural Network-GARCH Model for International Stock Return Volatility", *Journal of Empirical Finance*, 4 (1), 17-46.

[13] Drèze, J.H. (1976): "Bayesian Limited Information Analysis of the Simultaneous Equations Model", *Econometrica* 44, 1045-1075.

[14] Drèze, J.H. (1977): "'Bayesian Regression Analysis Using Poly-t Densities", *Journal of Econometrics* 6, 329-354.

[15] Edwards, R.G. and A.D. Sokal (1988), "Generalization of the Fortuin-Kasteleyn-Swendsen-Wang Representation and Monte Carlo Algorithm", *Physical Review D*, 38, 2009-2012

[16] Frühwirth-Schnatter (2004), S.: "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques", *Econometrics Journal* 7, 143-167.

[17] Gallant, A.R. and H. White (1989): "There exists a neural network that does not make avoidable mistakes", in *Proc. of the International Conference on Neural Networks*, San Diego, 1988 (IEEE Press, New York).

[18] Gallant, A.R. and G . Tauchen (1993): "A Nonparametric Approach to Nonlinear Time Series Analysis: Estimation and Simulation", in *New Directions in Time Series Analysis Part II*, ed. by D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, M.S. Taqqu. Springer-Verlag, New York.

[19] Gelman, A. and X. Meng (1991): "A Note on Bivariate Distributions That Are Conditionally Normal", *The American Statistician*, 45, 125-126.

[20] Geman, S. and D. Geman (1984): "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

[21] Geweke, J. (1989): "Bayesian inference in econometric models using Monte Carlo integration", *Econometrica*, 57, 1317-1339.

[22] Geweke, J. (1999): "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication", *Econometric Reviews*, 18(1), 1-73.

[23] Hammersley, J. and D. Handscomb (1964): *Monte Carlo Methods*. Chapman and Hall, London.

[24] Hastie, T., R. Tibshirani and J. Friedman (2001): *The Elements of Statistical Learning*, Springer-Verlag, New York.

[25] Hastings, W.K. (1970): "Monte Carlo Sampling Methods using Markov Chains and their Applications", *Biometrika*, 57, 97-109.

[26] Hecht-Nielsen, R. (1987): "Kolmogorov mapping neural network existence theorem", in *Proc. IEEE First International Conference on Neural Networks*, San Diego, 1987, 11-13.

[27] Hobert, J.P. and G. Casella (1996): "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models", *Journal of the American Statistical Association*, 91(436), 1461-1473.

[28] Hoogerheide, L.F. and H.K. van Dijk (2001): "Comparison of the Anderson-Rubin test for overidentification and the Johansen test for cointegration", Econometric Institute report 2001-04, Erasmus University Rotterdam.

[29] Hoogerheide, L.F., J.F. Kaashoek and H.K. van Dijk (2002): "Functional Approximations to Posterior Densities: A Neural Network Approach to Efficient Sampling", Econometric Institute report 2002-48, Erasmus University Rotterdam.

[30] Hoogerheide, L.F., J.F. Kaashoek and H.K. van Dijk (2004): "Neural network based approximations to posterior densities: a class of flexible sampling methods with applications to reduced rank models", Econometric Institute report 2004-19, Erasmus University Rotterdam.

[31] Hornik, K., M. Stinchcombe, and H. White (1989): "Multilayer feedforward networks are universal approximators", *Neural Networks*, Vol. 2, 359-366.

[32] Hutchinson, J., A. Lo and T. Poggio (1994): "A Nonparametric Approach to the Pricing and Hedging of Derivative Securities Via Learning Networks", *Journal of Finance*, 49, 851-889.

[33] Imbens, G.W. and J.D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects", *Econometrica* 62, 467-475

[34] Kleibergen, F.R., and H.K. Van Dijk (1994): "On the Shape of the Likelihood/Posterior in Cointegration Models", *Econometric Theory*, 10(3-4), 514-551.

[35] Kleibergen, F.R., and H.K. Van Dijk (1998): "Bayesian Simultaneous Equations Analysis using Reduced Rank Structures", *Econometric Theory*, 14(6), 701-743.

[36] Kloek, T., and H.K. Van Dijk (1978): "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo", *Econometrica*, 46, 1-19.

[37] Kolmogorov, A.N. (1957): "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition", *American Mathematical Monthly Translation*, Vol. 28, pp 55-59. (Russian original in Doklady Akademii Nauk SSSR, 144, 953-956)

[38] Leshno, M., Lin, V.Y., Pinkus, A. and Schocken, S. (1993): "Multilayer Feedforward Networks With a Nonpolynomial Activation Function Can Approximate Any Function", *Neural networks*, Vol. 6, 861-867.

[39] Liesenfeld, R. and J.-F. Richard (2002): "Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics", Discussion paper, University of Tubingen.

[40] Maddala, G.S. (1976): "Weak Priors and Sharp Posteriors in Simultaneous Equation Models", *Econometrica* 44, 345-351.

[41] Meng, X.-L. and W. H. Wong (1996): "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration", *Statistica Sinica* 6, 831-860.

[42] Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953): "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*, 21, 1087-1091.

[43] Paap, R. and H.K. van Dijk (2003): "Bayes Estimates of Markov Trends in Possibly Cointegrated Series: An Application to US Consumption and Income", *Journal of Business & Economic Statistics*, 21, 547-563.

[44] Pesaran, M.H. and R. Smith (1995): "Estimation of Long-Run Relationships from Dynamic Heterogeneous Panels", *Journal of Econometrics*, 68, 79-113.

[45] Richard, J.-F. (1998): "Efficient High-dimensional Monte Carlo Importance Sampling", Discussion paper, University of Pittsburgh.

[46] Ritter, C. and M.A. Tanner (1992): "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler", *Journal of the American Statistical Association*, 87, 861-868.

[47] Shephard, N. (1996): "Statistical aspects of ARCH and stochastic volatility", in *Time Series Models with Econometric, Finance and Other Applications*, ed. by D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielson, Chapman and Hall, London.

[48] Staiger, D. and J.H. Stock (1997): "Instrumental Variable Regression with Weak Instruments", *Econometrica*, 65, 557-586.

[49] Stinchcombe, M. (1989): "Universal Approximation Using Feedforward Networks with Non-sigmoid Hidden Layer Activation Functions", in *Proceedings of the International Joint Conference on Neural Networks, Washington DC*, IEEE Press, New York.

[50] Stinchcombe, M. (1990): "Approximating and Learning Unknown Mappings Using Multilayer Feedforward Networks with Bounded Weights", in *Proceedings of the International Joint Conference on Neural Networks, San Diego*, IEEE Press, New York.

[51] Strachan, R.W. and H.K. van Dijk (2004): "Improper priors with well defined Bayes Factors", Econometric Institute report 2004-18, Erasmus University Rotterdam.

[52] Tanner, M.A. and W.H. Wong (1987): "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528-550.

[53] Tierney, L. (1994): "Markov Chains for Exploring Posterior Distributions", *Annals of Statistics*, 22, 1701-1762.

[54] Van Dijk, H.K., and T. Kloek (1980): "Further experience in Bayesian analysis using Monte Carlo integration", *Journal of Econometrics*, 14, 307-328.

[55] Van Dijk, H.K., and T. Kloek (1984): "Experiments with some alternatives for simple importance sampling in Monte Carlo integration", in *Bayesian Statistics 2*, ed. by J. M. Bernardo, M. Degroot, D. Lindley, and A. F. M. Smith, Amsterdam, North-Holland.

[56] Van Dijk, H.K. (2003): "On Bayesian structural inference in a simultaneous equation model", in *Econometrics and the philosophy of economics*, ed. by B.P. Stigum, Princeton University Press, Princeton, New Jersey.

[57] Zellner, A. (1971): *An introduction to Bayesian inference in econometrics*. Wiley, New York.

[58] Zellner, A., L. Bauwens and H.K. van Dijk (1988): "Bayesian Specification Analysis and Estimation of Simultaneous Equation Models Using Monte Carlo Methods", *Journal of Econometrics* 38, 39-72.