Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 139 (2007) 154–180
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

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[a,*], Johan F. Kaashoek[b], Herman K. van Dijk[a,1]

[a]Econometric and Tinbergen Institutes, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands
[b]Econometric Institute, Erasmus University Rotterdam, The Netherlands

Available online 1 August 2006

Abstract

Likelihoods and posteriors of instrumental variable (IV) 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 posterior probabilities and marginal densities 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

[*]Corresponding author. Tel.: +31 10 4081424; fax: +31 10 4089162.
E-mail addresses: lhoogerheide@few.eur.nl (L.F. Hoogerheide), kaashoek@few.eur.nl (J.F. Kaashoek), hkvandijk@few.eur.nl (H.K. van Dijk).

[1]Part of this paper was written when the third author was visiting scholar at CORE, Université catholique de Louvain, Belgium.

doi:10.1016/j.jeconom.2006.06.009

approximates the target density. The methods are tested on a set of illustrative IV regression models. The results indicate the possible usefulness of the neural network approach.

JEL classification: C11; C15; C45

Keywords: Instrumental variables; Reduced rank; Importance sampling; Markov chain Monte Carlo; Neural networks

1. Introduction

There exist classes of statistical and econometric models where the conditional distribution of any parameter of interest, given the other parameters, has known analytical properties and elliptically shaped Bayesian HPD credible sets, see e.g. Berger (1985). However, the joint and marginal distributions of the parameters may have unknown analytical properties and non-elliptical HPD credible sets. Then 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 econometric models we refer to e.g. Imbens and Angrist (1994) and Bos et al. (2000) and the references cited there. A canonical statistical model is given by Gelman and Meng (1991). A second issue is that one may have great difficulties when trying to simulate (pseudo-) random drawings from such a class of non-elliptical joint distributions; random drawings are required for inference on nonlinear functions of parameters of interest such as impulse responses, see Strachan and Van Dijk (2004). Even if it is relatively easy to simulate random drawings from the conditional distributions, multi-modality and/or high correlations may cause the Gibbs sampler to converge extremely slowly or even yield erroneous results.

A first contribution of this paper is to show that well-behaved conditional distributions of parameters of interest may occur together with ill-behaved marginals for the case of linear models with reduced rank. We focus on the class of instrumental variable (IV) regression models with possibly endogenous regressors. This class of models may exhibit reduced rank of the parameter matrix due to varying degrees of instrument quality and endogeneity. Under certain weak priors the conditional posterior distributions in this model are Student's *t*, that is, at least if they are proper. In the presence of weak instruments the joint and marginal posteriors may, however, display highly non-elliptical contours.

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. In the first step a neural network is constructed that approximates the target density. In the second step this neural network is embedded in a

Metropolis–Hastings (MH) or importance sampling (IS) algorithm.[2] 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.[3] 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 unimodal. If the target (posterior) distribution is multimodal 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.

The proposed methods are applied on a set of illustrative examples of posterior distributions in IV regression models. 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.[4]

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 HPD 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 illustrates the neural network approach in examples of IV regressions with simulated data. Conclusions are given in Section 5.

2. On the shape of posterior densities and Bayesian credible sets in IV regression models with several degrees of endogeneity and instrument quality

In this section we analyze a class of models, IV regression models with possibly endogenous regressors, where the conditional posterior distributions of parameters of interest have known properties but the joint does not. Consider the following possibly overidentified IV model, also known as the incomplete simultaneous equations model (INSEM). Following Zellner et al. (1988), let:

$$y_1 = y_2\beta + \varepsilon, \tag{1}$$

$$y_2 = X\pi + v, \tag{2}$$

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

[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). IS, 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).

[3] Kolmogorov (1957) and Hecht-Nielsen (1987) establish general theoretical capabilities. Proofs concerning neural network approximations for specific configurations can be found in Gallant and White (1988), Hornik et al. (1989) and Leshno et al. (1993).

[4] 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.

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. Assume that the rows of the matrix of error terms $(\varepsilon\ v)$ are independently normally distributed with (2×2) covariance matrix Σ with elements σ_{ij} $(i,j = 1,2)$. A well-known example is the stylized wage regression where y_1 is the log of hourly wage and y_2 denotes education which is possibly endogenous owing to the omission of a variable measuring (unobservable) ability. 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 an IV. Staiger and Stock (1997) show that classical inference on the rate of return to schooling, β, can be greatly affected by the weakness of the quarter of birth instruments.

We specify the following non-informative prior density:

$$p(\beta, \pi, \Sigma) \propto |\Sigma|^{-h/2} \quad \text{with } h > 0. \tag{3}$$

Given the model (1)–(3), one can easily derive the likelihood function and the posterior density kernel of (β, π, Σ). Using properties of the inverted Wishart distribution, see Zellner (1971) and Bauwens and Van Dijk (1990), in order to integrate Σ out of the joint posterior, and choosing $h = 3$ in the prior density kernel (3) leads to the following joint posterior kernel of (β, π):

$$p(\beta, \pi | y_1, y_2, X) \propto \begin{vmatrix} (y_1 - y_2\beta)'(y_1 - y_2\beta) & (y_1 - y_2\beta)'(y_2 - X\pi) \\ (y_2 - X\pi)'(y_1 - y_2\beta) & (y_2 - X\pi)'(y_2 - X\pi) \end{vmatrix}^{-T/2}. \tag{4}$$

For $\pi = 0$ the posterior kernel in (4) reduces to the (non-zero) 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. For $\pi \neq 0$ the integral $\int p(\beta, \pi | y_1, y_2, X)\,\mathrm{d}\beta$ is finite; however, when $\pi \to 0$ this integral increases at a rate of $(\pi' X' M_{y_2} X\pi)^{-1/2}$, so that $\iint p(\beta, \pi | y_1, y_2, X)\,\mathrm{d}\beta\,\mathrm{d}\pi$ is not finite. For details, see Propositions 3 and 4. So, the joint density of β and π is improper on $\mathbb{R}^{k+1}$. Although improper on $\mathbb{R}^{k+1}$, the posterior in (4) can be made proper by restricting β and/or π to a certain area. In that case it depends on the data y_1, y_2 and X, whether the behavior for $\pi = 0$ still dominates the analysis.

For illustrative purposes, the posterior kernel in (4) is calculated for simulated data sets from (1) to (2) with $k = 1$, $T = 100$, $\beta = 0$, $\sigma_{11} = \sigma_{22} = 1$ for nine cases. In each case we use the same vector of instruments denoted by x, where the elements of x are i.i.d. N(0,1) drawings. 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/strong 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. Fig. 1 shows contour plots of the joint posterior kernel of β and π in (4) 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, bimodality and elongated ridges. Table 1 gives an overview of the possible shapes of the joint posterior kernel of β

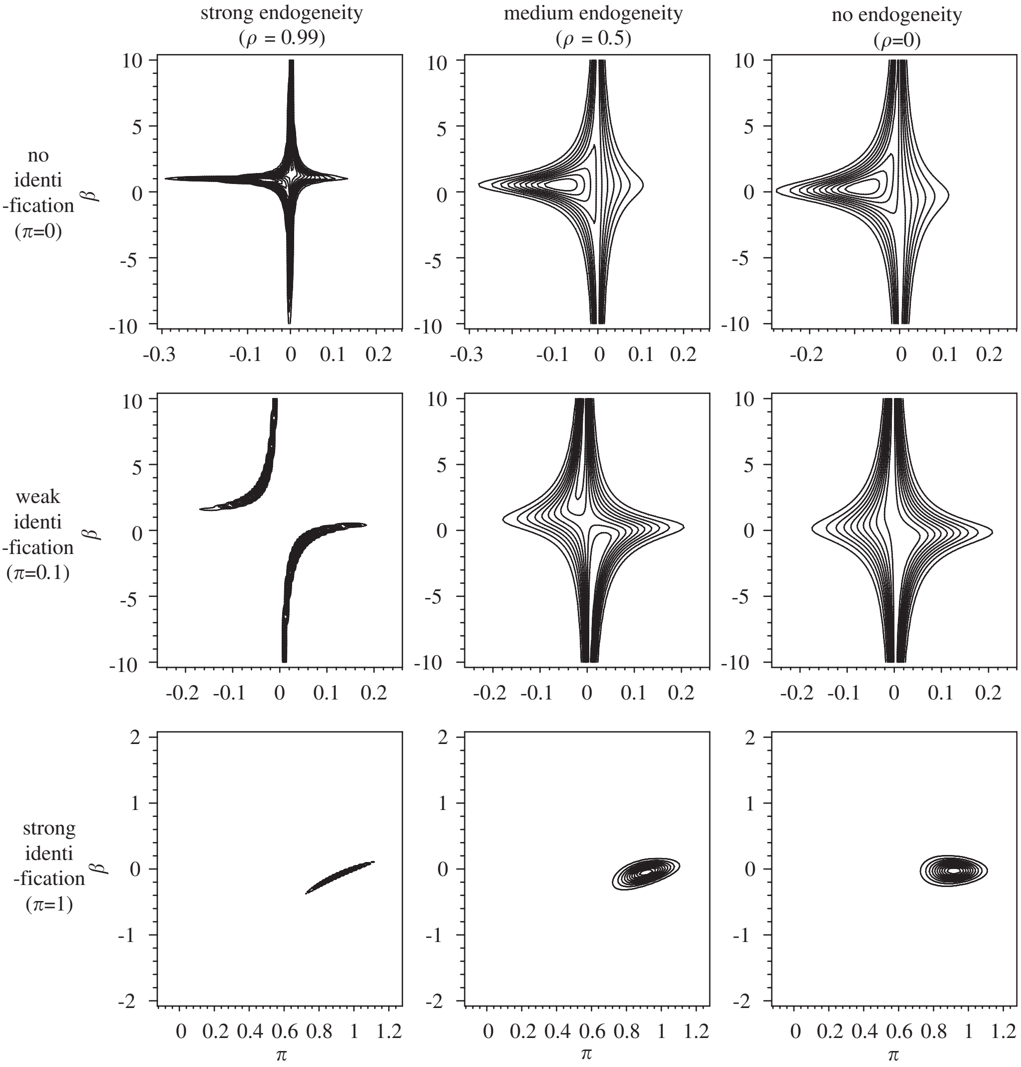

Fig. 1. Contour plots in the $\pi \times \beta$ plane: joint posterior kernel of π and β in (4) in IV model for nine simulated data sets; three cases of identification ($\pi = 0$, 0.1, 1 corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99$, 0.5, 0 corresponding to strong, medium, no endogeneity).

and π in this simple IV regression model with $k = 1$ instrument for different cases of simulated data.[5]

Note that in the three cases of simulated data sets with strong instruments ($\pi = 1$), the contour plots do not show a high-level ridge at $\pi = 0$; the value of the joint posterior kernel

[5] We have repeated the experiment 10 times with different seeds of the random number generator. In five of the nine cases bimodality showed up in the contour plot in two simulations; this is indicated as ‘possibly bimodality’. Repeating the simulation with a different value of β yields the same shapes. For some related graphs we refer to Van Dijk (2003).

Table 1
Shape of the posterior density kernel of β and π in the IV regression model (1)–(2) with one instrument and weak prior (3) 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 and possibly bimodality
	Weak	Ridges and bimodality	Ridges and possibly bimodality	Ridges and possibly bimodality
	Strong	Nearly elliptical	Elliptical	Elliptical

for $\pi = 0$ is relatively very small as compared to the value of the joint posterior kernel at its mode. In the just identified model (with $k = 1$) the mode is given by ($\hat{\beta}_{2SLS} = y_1'x/y_2'x, \hat{\pi}_{OLS} = y_2'x/x'x$), and the ratio between the posterior kernel in (4) for $\pi = 0$ (and arbitrary β) and the value at its mode ($\hat{\beta}_{2SLS}, \hat{\pi}_{OLS}$) is

$$\frac{p(\beta, \pi = 0|y_1, y_2, X)}{p(\hat{\beta}_{2SLS}, \hat{\pi}_{OLS}|y_1, y_2, X)} = \left[1 - \frac{r^2_{y_2,x} + r^2_{y_1,x} - 2r_{y_1,x}r_{y_2,x}r_{y_1,y_2}}{1 - r^2_{y_1,y_2}}\right]^{T/2}, \tag{5}$$

where $r_{y_2,x} \equiv y_2'x/\sqrt{y_2'y_2\ x'x}$, etc. In the three cases of strong instruments (with large $r^2_{y_2,x}$) as well as in the case of weak instruments and strong endogeneity (with $r^2_{y_1,y_2}$ close to one) the ratio (5) is small ($<10^{-9}$). Also in the case of irrelevant instruments and strong endogeneity $1 - r^2_{y_1,y_2}$ is small; however, as $r_{y_1,x}$ and $r_{y_2,x}$ are both small and $r_{y_1,x} \approx r_{y_2,x}$, the numerator on the right-hand side of (5) is even much smaller, so that in this case the contour plot displays a high-level ridge at $\pi = 0$.

If we consider the contour plot of the posterior kernel (4) raised to the power $\frac{1}{20}$, so that the contour plot also shows the contours for much lower values of the posterior kernel, we observe also in the case of strong identification the presence of bimodality or an elongated ridge around the line $\pi = 0$; see Fig. 2. The origin of these hyperbolic contour lines becomes intuitively clear if we consider the fact that the structural form (1)–(2) is equivalent with the orthogonal structural form (see Zellner et al., 1988):

$$y_1 = y_2\beta + v\phi + \eta, \tag{6}$$

$$y_2 = X\pi + v, \tag{7}$$

where $\phi = \sigma_{12}/\sigma_{22}$; η and v are mutually independent, i.i.d. Gaussian error terms. Eq. (6) is equivalent with

$$y_1 = y_2\gamma_1 + X\gamma_2 + \eta, \tag{8}$$

where $\gamma_1 = \beta + \phi$, $\gamma_2 = -\pi\phi$, so that $\gamma_2 = \pi(\beta - \gamma_1)$, and in the case of $k = 1$ instrument $\beta = \gamma_1 + \gamma_2/\pi$. In the just identified model the set of points (β, π) for which the posterior kernel in (4) scaled by the value at its mode, $p(\hat{\beta}_{2SLS}, \hat{\pi}_{OLS}|y_1, y_2, X)$, has a certain value $C \in (0, 1]$ is given by

$$\{\beta = \hat{\gamma}_1 + \hat{\gamma}_2/\pi \pm \sqrt{p_C(\pi)}/\pi, p_C(\pi) \geqslant 0, \pi \neq 0\}, \tag{9}$$

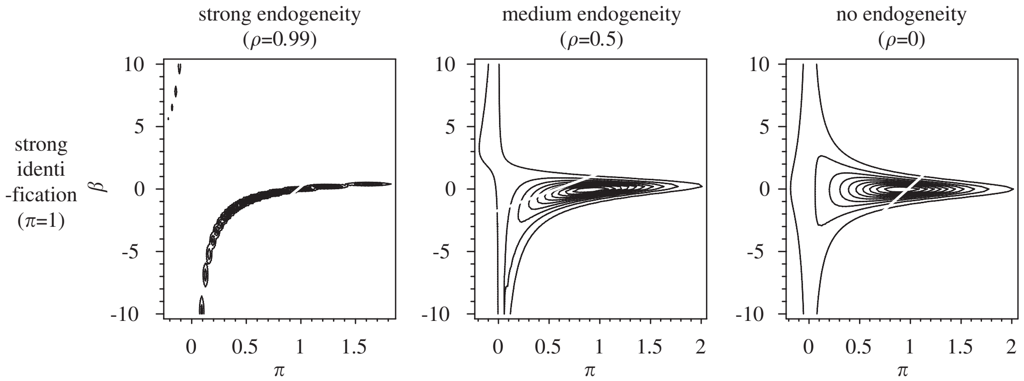

Fig. 2. Contour plots in the $\pi \times \beta$ plane: joint posterior kernel of β and π in (4) raised to the power 1/20 in IV model for three simulated data sets; the case of strong identification ($\pi = 1$) combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity).

where $(\hat{\gamma}_1, \hat{\gamma}_2)$ are the OLS estimators in (8), and $p_C(\pi)$ is a polynomial of degree 2 that is non-negative on the interval with bounds $\hat{\pi}_{\text{OLS}} \pm (s_{y_2}/s_x)\sqrt{(1 - r^2_{y_2,x})(C^{-2/T} - 1)}$ with $s_{y_2} \equiv \sqrt{y_2'y_2}$, which includes $\pi = 0$ for C small enough. In the three cases of a strong instrument $\hat{\pi}_{\text{OLS}}$ is far from zero (with t-value of $\hat{\pi}_{\text{OLS}}$ larger than 10), resulting in (nearly) elliptical shapes far away from $\pi = 0$. In the cases of no/weak identification $\hat{\pi}_{\text{OLS}}$ is small (with t-value smaller than 1). In these cases the shapes depend on $\hat{\gamma}_2$: if (the t-value of) $\hat{\gamma}_2$ is close to zero, the contour plot shows connected ridges around $\pi = 0$; otherwise it displays two disconnected ridges on both sides of $\pi = 0$ (and on both sides of $\beta = \hat{\gamma}_1$where $\hat{\gamma}_1 \approx 1$ in the presence of strong endogeneity). The squared t-value of $\hat{\gamma}_2$ is equal to

$$t^2_{\hat{\gamma}_2} = (T-2)\left(\frac{(1 - r^2_{y_1,y_2})(1 - r^2_{y_2,x})}{(r_{y_1,x} - r_{y_2,x}r_{y_1,y_2})^2} - 1\right)^{-1},$$

which is large in the case of weak identification and strong endogeneity as $(1 - r^2_{y_1,y_2})$ is small and the weak influence of x on y_2 causes a certain difference between $r_{y_1,x}$ and $r_{y_2,x}$.

As can be seen from Fig. 2 and formula (9), even in the presence of strong 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.

So, the posterior density kernel of β and π may show highly non-elliptical shapes if instruments are weak. Drèze (1976, 1977) and Kleibergen and Van Dijk (1994b, 1998) present theoretical results on the conditional and marginal distributions of β and π corresponding to this joint density kernel. We reformulate, extend and illustrate their results for the simple IV regression model (1)–(3).

2.1. Weak and strong structural inference

In Drèze (1976, 1977) the conditional posterior density of β given π and the marginal posterior density of β are derived. We summarize and reformulate his results in two propositions:

Proposition 1 (*Conditional posterior of β given π*). *In the IV regression model* (1)–(2) *with prior* (3) *the conditional posterior density of β given π* (*with $\pi \neq 0$*) *is a Student's 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; $M_v \equiv I - v(v'v)^{-1}v'$ with $v \equiv y_2 - X\pi$, i.e. v is a function of the parameter π (*and the data y_2, X*) *instead of the vector of simulated error terms.*

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 is improper. For $\pi \neq 0$ conditional HPD credible sets of β are elliptical; in this case the conditional mean is equal to the OLS estimator of β in the orthogonal structural form Eq. (6).

Proposition 2 (*Marginal posterior of β*). *In the IV regression model* (1)–(2) *with prior* (3) *the marginal posterior density of β is proportional to the ratio of two Student's t kernels*:

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

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 $(k-1)$. The marginal posterior of β tends to a bell-shaped function as long as the number of instruments k becomes large enough, which 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 *number* of instruments still yields a bell-shaped marginal posterior of β. This result appeared in an informal way in Maddala (1976), commenting on Drèze (1976).

Fig. 3 shows the marginal posterior of β in (11) for our nine simulated data sets; the posterior kernels are normalized over the displayed range. Notice that the graphs display fat tails in the cases of no identification, combined with a sharp peak in the case of strong endogeneity; in these cases the kernel (11) is approximately equal to $[(y_1 - y_2\beta)'(y_1 - y_2\beta)]^{-k/2}$ as $M_X y_1 \approx y_1, M_X y_2 \approx y_2$. Also note the bimodality in the case of the weak instrument and strong endogeneity; this results from the term

$$\left[\frac{(y_1 - y_2\beta)' M_X (y_1 - y_2\beta)}{(y_1 - y_2\beta)'(y_1 - y_2\beta)} \right]^{(T-k-1)/2} = \left[1 - \frac{y_2' P_X y_2 \, (\beta - \hat{\beta}_{2\text{SLS}})^2}{y_1' M_{y_2} y_1 + y_2' y_2 (\beta - \hat{\beta}_{\text{OLS}})^2} \right]^{(T-k-1)/2} \tag{12}$$

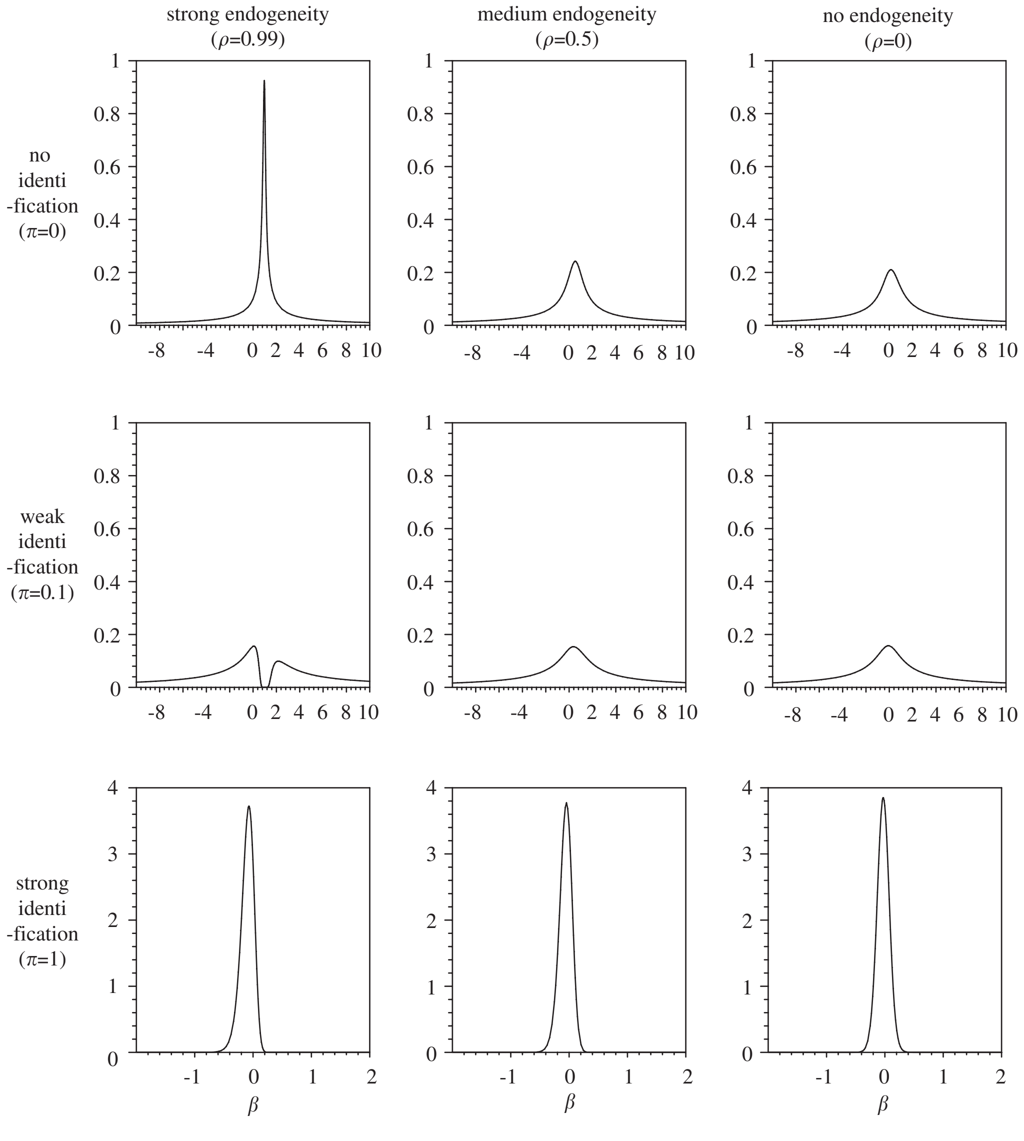

Fig. 3. Marginal posterior kernel of β in (11) in IV model for nine simulated data sets; three cases of identification ($\pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity).

with $P_X \equiv X(X'X)^{-1}X'$, which is equal to

$$\left[1-\frac{r^2_{y_2',x}\left(\frac{\beta-\hat{\beta}_{2\mathrm{SLS}}}{s_{y_1}/s_{y_2}}\right)^2}{1-r^2_{y_1,y_2}+\left(\frac{\beta-\hat{\beta}_{\mathrm{OLS}}}{s_{y_1}/s_{y_2}}\right)^2}\right]^{(T-2)/2} \tag{13}$$

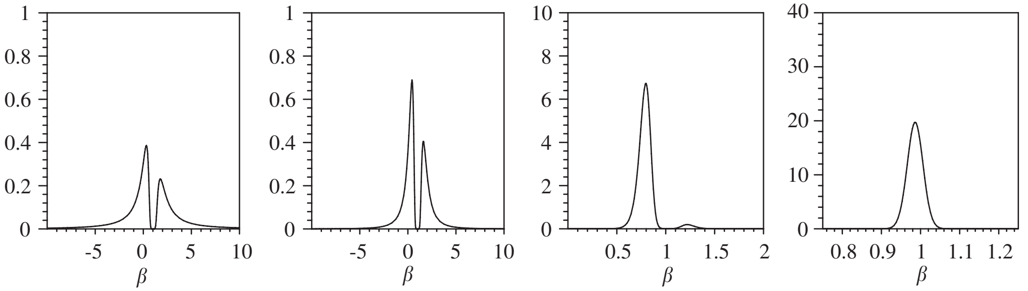

Fig. 4. Marginal posterior kernel of β in (11) in IV model for simulated data with weak identification ($\pi = 0.1$) and strong endogeneity ($\rho = 0.99$) after adding 1, 2, 15 or 75 irrelevant (i.i.d. N(0,1)) instruments, respectively.

with $s_{y_i} \equiv \sqrt{y_i'y_i}$ $(i = 1, 2)$ in the case of $k = 1$ instrument. In the case of one weak instrument and strong endogeneity $\hat{\beta}_{2\text{SLS}}$ and $\hat{\beta}_{\text{OLS}}$ are in general far apart, while $r^2_{y_2,x}$ is small and $r^2_{y_1,y_2}$ is close to one, so that (13) takes very small values near $\beta = \hat{\beta}_{\text{OLS}} \approx 1$, whereas on both sides of $\hat{\beta}_{\text{OLS}}$ there is an interval where (13) is not negligible. In the cases with a strong instrument the graphs show a bell shape; in these cases the term (13), converging to the very small constant $(1 - r^2_{y_2,x})^{(T-2)/2}$ when β becomes large (in absolute sense), makes the graph seem to be bell-shaped; also in these cases (13) is very small near $\beta = \hat{\beta}_{\text{OLS}}$ if $\hat{\beta}_{2\text{SLS}}$ and $\hat{\beta}_{\text{OLS}}$ are far apart, but the large value of $r^2_{y_2,x}$ causes (13) to be only large on one relatively small interval around $\beta = \hat{\beta}_{2\text{SLS}}$, so that the graphs do not display bimodality. For a more detailed analysis comparing Bayesian and classical inference in an IV regression model, we refer to Kleibergen and Zivot (2003).

In Fig. 4 the marginal posterior kernel of β is shown where 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. However, notice that the location of the bell shape in the case of many irrelevant instruments is different from the case of a strong instrument: many irrelevant instruments yield a bell shape around $\hat{\beta}_{\text{OLS}}$, which is far away from the true value of $\beta = 0$ in this case of strong endogeneity, whereas a strong instrument yields a bell shape in the neighborhood of $\beta = 0$.

2.2. *Impossible restricted reduced form inference*

In Kleibergen and Van Dijk (1994b, 1998) the conditional posterior density of π given β and the marginal posterior density kernel of π are derived. We summarize their results in two propositions:

Proposition 3 (*Conditional posterior of π given β*). *In the IV regression model* (1)–(2) *with prior* (3) *the conditional posterior density of π given β is a Student's 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*:

$$p(\pi|\beta, y_1, y_2, X) = c_2|s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1}|^{-1/2} \times \left[1 + \frac{1}{T-k}(\pi - \hat{\pi})'(s^2_{\hat{\pi}}(X'M_\varepsilon X)^{-1})^{-1}(\pi - \hat{\pi})\right]^{-T/2}, \qquad (14)$$

where $(T-k)s_{\hat{\pi}}^2 \equiv (y_2 - X\hat{\pi})' M_\varepsilon (y_2 - X\hat{\pi})$ *and* c_2 *is a scaling constant that only depends on* T *and* k; $M_\varepsilon \equiv I - \varepsilon(\varepsilon'\varepsilon)^{-1}\varepsilon'$, *with* $\varepsilon \equiv y_1 - y_2\beta$.

For all values of β this density exists. HPD credible sets are elliptical.

Proposition 4 (*Marginal posterior of* π). *In the IV regression model* (1)–(2) *with prior* (3) *the marginal posterior density of* π *is proportional to the ratio of a product of two Student's t kernels in the numerator and one Student's t kernel in the denominator*:

$$p(\pi|y_1, y_2, X) \propto \frac{[(y_2 - X\pi)'(y_2 - X\pi)]^{-(T-1)/2}(\pi'X'M_{[y_1 y_2]}X\pi)^{-(T-1)/2}}{(\pi'X'M_{y_2}X\pi)^{-(T-2)/2}} \tag{15}$$

$$= [(y_2 - X\pi)'(y_2 - X\pi)]^{-(T-1)/2} \times (\pi'X'M_{y_2}X\pi)^{-1/2}\left(\frac{\pi'X'M_{y_2}X\pi}{\pi'X'M_{[y_1 y_2]}X\pi}\right)^{(T-1)/2}, \tag{16}$$

known as the 2–1 *poly t density*.

The density kernel is not integrable over neighborhoods around zero (because of the term $(\pi'X'M_{y_2}X\pi)^{-1/2}$), so that this is not a proper density. Given this non-integrability, reduced form inference on π is not possible. This result does *not* depend on the quality or quantity of the instruments *nor* on the endogeneity in the data. Only if the restriction that y_2 is not an endogenous regressor, $\sigma_{12} = 0$, is imposed on the model *beforehand* we obtain a proper marginal density of π. For example, specifying $p(\beta, \pi, \sigma_{11}, \sigma_{22}) \propto \sigma_{11}^{-1/2}\sigma_{22}^{-1/2}$ 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 given by $p(\beta, \pi|y_1, y_2, X) \propto [(y_1 - y_2\beta)'(y_1 - y_2\beta)]^{-T/2}[(y_2 - X\pi)'(y_2 - X\pi)]^{-T/2}$, i.e. β and π have independent Student's t distributions with $T-1$ and $T-k$ degrees of freedom, respectively.

So, in model (1)–(3) forecasting future values of y_2 using posterior moments of π is not possible if one uses 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 $\pi = 0$. Fig. 5 shows the marginal posterior density kernel of π in (16) for our nine simulated data sets. Note 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 around $\pi = \hat{\pi}_{\text{OLS}}$ (≈ 0.9 for this simulated data set).

It may seem paradoxical that if Eq. (1) is excluded from the model, forecasting based on (2) is standard, whereas adding the extra information in Eq. (1), $y_1 = y_2\beta + \varepsilon$ with ε possibly correlated with v, makes this impossible. However, as Kleibergen and Van Dijk (1994a) and Chao and Phillips (1998) point out, the flat prior for (β, π) implies a highly informative prior for the parameters (π_1, π) of the restricted reduced form

$$y_1 = X\pi_1 + v_1, \tag{17}$$

$$y_2 = X\pi + v, \tag{18}$$

where $\pi_1 = \pi\beta$ and $v_1 = v\beta + \varepsilon$; in the just identified model ($k = 1$) there exists a 1–1 relationship between (β, π) and (π_1, π), so that in that case it is easily derived that

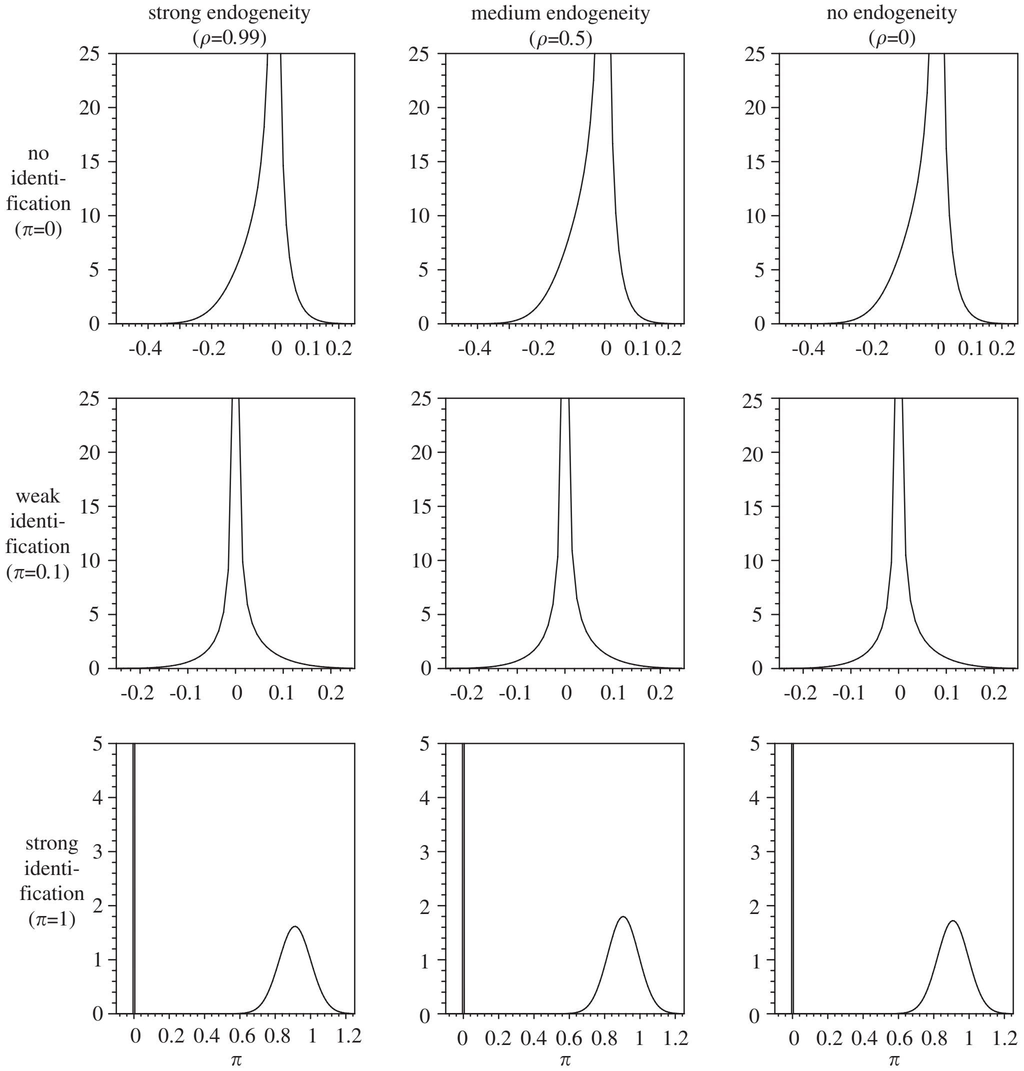

Fig. 5. Marginal posterior kernel of π in (15) in IV model for nine simulated data sets; three cases of identification ($\pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity). An asymptote at $\pi = 0$ occurs in each figure.

$p(\pi_1, \pi) \propto p(\beta, \pi)|\partial(\beta, \pi)/\partial(\pi_1, \pi)| = |\pi|^{-1}$: the prior for (π_1, π) is far from non-informative for π, as it gives infinite density to the point $\pi = 0$.

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 drawings from the density (kernel) $nn(\theta)$.
3. Perform IS or the (independence chain) MH 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:

$$nn(\theta) = eG_2(CG_1(A\theta + b) + d) + f, \quad \theta \in \mathbb{R}^n, \tag{19}$$

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 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

$$G_1(v) = (g_1(v_1), \ldots, g_1(v_{H_1}))', \quad G_2(z) = (g_2(z_1), \ldots, g_2(z_{H_2}))', \;\; v \in \mathbb{R}^{H_1}, \;\; z \in \mathbb{R}^{H_2}, \tag{20}$$

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

The following three specifications of (19) allow for easy sampling (when this neural network function is considered as a density kernel):

Type 1 *neural network*: A standard three-layer feed-forward neural network (in the notation of (19): $H_2 = 1$, $e = 1$, $f = 0$ and g_2 is the identity $g_2(x) = x$, $x \in \mathbb{R}$). As activation function g_1 in (20) we take the scaled arctangent function:

$$g_1(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}, \quad x \in \mathbb{R}. \tag{21}$$

The reason for choosing the arctangent function is that it can be analytically integrated infinitely many times. We show in Section 3.2, that this property makes the neural network, in the role of a density kernel on a bounded region, easy to sample from. The scaling is merely done because it is common practice to use activation functions that take values in the unit interval.

Type 2 *neural network*: A simplified four-layer network with the second hidden layer consisting of only one cell ($H_2 = 1$, $e = 1$, $f = 0$), g_2 the exponential function and activation function g_1 in (20) equal to the following piecewise-linear function *plin*:

$$plin(x) = \begin{cases} 0, & x < -1/2, \\ x + 1/2, & -1/2 \leqslant x \leqslant 1/2, \\ 1, & x > 1/2, \end{cases} \quad x \in \mathbb{R}. \tag{22}$$

We show in Section 3.2 that these activation functions make 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 Section 3.2. Another example of such a function is the logistic function.

Type 3 *neural network*: A mixture of Student's *t* distributions:

$$nn(\theta) = \sum_{h=1}^{H} p_h t(\theta|\mu_h, \Sigma_h, v), \tag{23}$$

where p_h $(h = 1, \ldots, H)$ are the probabilities of the Student's *t* components and where $t(\theta|\mu_h, \Sigma_h, v)$ is a multivariate *t* density with mode vector μ_h, scaling matrix Σ_h, and v degrees of freedom:

$$t(\theta|\mu_h, \Sigma_h, v) = \frac{\Gamma((v+n)/2)}{\Gamma(v/2)(\pi v)^{n/2}} |\Sigma_h|^{-1/2} \left(1 + \frac{(\theta-\mu_h)'\Sigma_h^{-1}(\theta-\mu_h)}{v}\right)^{-(v+n)/2}. \tag{24}$$

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 (19), $H_2 = H$ (the number of *t* densities), $H_1 = Hn$, activation functions

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

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

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

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's *t* distribution has fatter tails than the normal distribution. Note that the p_h $(h = 1, \ldots, H)$ in (23) have to satisfy $\sum_{h=1}^{H} p_h = 1$. Because of this restriction the approximation capabilities do not directly follow from the references cited in footnote 3. However, Zeevi and Meir (1997) show that under certain conditions any density function may be approximated to arbitrary accuracy by a convex combination of 'basis' densities; the mixture of Student's *t* densities in (23) falls within their framework.

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 et al. (2001), such methods are termed 'adaptive basis function methods' or 'dictionary methods'. A key ingredient of these methods is a search mechanism that constructs a linear combination of (nonlinear) basis functions that are chosen from a (possibly infinite) set or 'dictionary' of candidate basis functions.

3.1. Constructing a neural network approximation to a density

3.1.1. 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, obtain a set of drawings

θ^i $(i = 1, \ldots, N)$ from the uniform distribution on the bounded region to which we restrict the random variable $\theta \in \mathbb{R}^n$ to take its values. Then approximate the target density kernel $p(\theta)$ with a neural network by minimizing the sum of squared residuals:

$$SSR(A, b, c, d) = \sum_{i=1}^{N} (p(\theta^i) - nn(\theta^i | A, b, c, d))^2, \tag{25}$$

where the notation c instead of C is used, since in our Type 1 and 2 networks this is a $(1 \times H_1)$ vector. We choose the most parsimonious 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 $(i = 1, \ldots, N)$.

Next, check the squared correlation R^2 between nn and p for a larger set of points than the 'estimation set'. If this R^2 is also high enough, then we may conclude that the network does not only provide a good approximation to p in the points θ^i $(i = 1, \ldots, N)$ but also in between, so that the approximation is really accurate. Otherwise, increase the number of points N and start all over again; for example, 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 a 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 establish this a penalty term is added to (25), for example $-M\sum_{i=1}^{N} I\{nn\ (\theta^i) < 0\} nn(\theta^i)$ where M is a constant large enough to make nn positive (or only slightly negative) in all points θ^i $(i = 1, \ldots, N)$. Notice that if the minimum of $nn(\theta)$ is an (in absolute sense) very small negative value, one can simply subtract this negative value from the network's constant d, so that $nn(\theta)$ becomes non-negative for each θ. 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. In our Type 2 (simplified four-layer) neural network the exponential function implies that non-negativity is automatically taken care of.

3.1.2. 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, compute the mode μ_1 and scale Σ_1 of the first Student's t distribution in the mixture as $\mu_1 = \text{argmax}_\theta\, p(\theta)$, the mode of the target distribution, and Σ_1 as minus the inverse Hessian of $\log p(\theta)$ evaluated at its mode μ_1. Then draw a set of points θ^i $(i = 1, \ldots, N)$ from the 'first stage neural network' $nn(\theta) = t(\theta | \mu_1, \Sigma_1, v)$, with small v to allow for fat tails.[6] After that add components to the mixture, iteratively, by performing the following steps:

[6] Throughout this paper we use Student's t distributions with $v = 1$. 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's t distributions and/or allow for different degrees of freedom in different Student's t distributions. This is a topic for further research.

Step 1: Compute the IS weights $w(\theta^i) = p(\theta^i)/nn(\theta^i)$ $(i = 1, \ldots, N)$. In order to determine the number of components H of the mixture we make use of a simple diagnostic criterion: the coefficient of variation, i.e. 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 IS weights caused by adding one new Student's t component to the candidate mixture is small, e.g. less than 10%, then stop: the current $nn(\theta)$ is the Type 3 neural network approximation.[7] Otherwise, go to step 2.

Step 2: Add another Student's t distribution with density $t(\theta|\mu_h, \Sigma_h, v)$ to the mixture with $\mu_h = \text{argmax}_\theta\, w(\theta) = \text{argmax}_\theta\{p(\theta)/nn(\theta)\}$ and Σ_h equal to minus the inverse Hessian of $\log w(\theta) = \log p(\theta) - \log nn(\theta)$ evaluated at its mode μ_h. Here, $nn(\theta)$ denotes the mixture of $(h-1)$ Student's t densities obtained in the previous iteration of the procedure. An obvious initial value for the maximization procedure for computing $\mu_h = \text{argmax}_\theta\, w(\theta)$ 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 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 minus the 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:

$$res(\theta) = \max\{p(\theta) - \tilde{c}\; nn(\theta), 0\}, \tag{26}$$

where $\tilde{c}$ 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 IS 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 $w_{\text{res}}(\theta^i)$ and scaled weights $\tilde{w}_{\text{res}}(\theta^i)$ $(i = 1, \ldots, N)$ are:

$$w_{\text{res}}(\theta^i) = \frac{res(\theta^i)}{nn(\theta^i)} = \max\{w(\theta^i) - \tilde{c}, 0\} \quad \text{and} \quad \tilde{w}_{\text{res}}(\theta^i) = \frac{w_{\text{res}}(\theta^i)}{\sum_{i=1}^{N} w_{\text{res}}(\theta^i)}, \tag{27}$$

and μ_h and Σ_h are obtained as:

$$\mu_h = \sum_{i=1}^{N} \tilde{w}_{\text{res}}(\theta^i)\theta^i, \quad \Sigma_h = \sum_{i=1}^{N} \tilde{w}_{\text{res}}(\theta^i)(\theta^i - \mu_h)(\theta^i - \mu_h)'. \tag{28}$$

There are two issues relevant for the choice of $\tilde{c}$ in (26) and (27). 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 $w(\theta^i) > \tilde{c}$ in order to make Σ_h non-singular. A procedure is to calculate Σ_h for $\tilde{c}$ equal to 100 times the average value of $w(\theta^i)$ $(i = 1, \ldots, N)$; if Σ_h in (28) is non-singular, accept $\tilde{c}$; otherwise lower $\tilde{c}$.

[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 $p(\theta)/nn(\theta)$ has relatively little variation.

Step 3: Choose the probabilities p_h $(h = 1, \dots, H)$ in the mixture $nn(\theta) = \sum_{h=1}^{H} p_h t(\theta|\mu_h, \Sigma_h, v)$ by minimizing the (squared) coefficient of variation of the IS weights. First, draw N points θ_h^i from each component $t(\theta|\mu_h, \Sigma_h, v)$ $(h = 1, \dots, H)$. Then minimize $\mathrm{E}[w(\theta)^2]/\mathrm{E}[w(\theta)]^2$, where:

$$\mathrm{E}[w(\theta)^k] = \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H} p_h w(\theta_h^i)^k \quad (k = 1, 2), \quad w(\theta_h^i) = \frac{p(\theta_h^i)}{\sum_{l=1}^{H} p_l t(\theta_h^i|\mu_l, \Sigma_l, v)}. \tag{29}$$

Step 4: Draw a sample of N points θ^i $(i = 1, \dots, N)$ from our new mixture of t distributions, $nn(\theta) = \sum_{h=1}^{H} p_h t(\theta|\mu_h, \Sigma_h, v)$, and go to step 1; in order to draw a point from the density $nn(\theta)$ first use a drawing from the U(0, 1) distribution to determine which component $t(\theta|\mu_h, \Sigma_h, v)$ is chosen, and then draw from this multivariate t distribution.

It may occur that one is dissatisfied with diagnostics like 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 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*

3.2.1. *Type 1 (3-layer) neural network density*

Suppose the joint density kernel of a certain $\theta \in \mathbb{R}^n$ is given by our Type 1 neural network:

$$nn(\theta) = \sum_{h=1}^{H} \frac{c_h}{\pi} \arctan(a_h'\theta + b_h) + \frac{1}{2}\sum_{h=1}^{H} c_h + d, \tag{30}$$

where each element θ_j is restricted to a certain finite interval $[\underline{\theta}_j, \bar{\theta}_j]$ $(j = 1, \dots, n)$. The arctangent is analytically integrable infinitely many times; its integrals are given by Theorem 1:

Theorem 1. *The nth integral* $J_n(x)$ $(n = 1, 2, \dots)$ *of the arctangent function*, $J_n(x) \equiv \int \cdots \int \arctan(x)\, \mathrm{d}x \cdots \mathrm{d}x$ *with* $x \in \mathbb{R}$, *is given by*

$$J_n(x) = p_n(x) \arctan(x) + q_n(x) \ln(1 + x^2) + r_n(x), \quad x \in \mathbb{R}, \tag{31}$$

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

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

with coefficients $p_{n,k}$ $(k = 0, 1, \dots, n)$ *and* $q_{n,k}$ $(k = 0, 1, \dots, n - 1)$ *given by*

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

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

Proof. By induction; see Hoogerheide et al. (2004).[8]

A kernel of the cumulative distribution function of $\theta \sim nn(\theta)$ with nn in (30) is given by

$$CDF_{\theta}(\theta_1, \ldots, \theta_n) = \left(\frac{1}{2}\sum_{h=1}^{H} c_h + d\right)(\theta_1 - \underline{\theta}_1)\ldots(\theta_n - \underline{\theta}_n) + \sum_{h=1}^{H} \frac{c_h}{\pi a_{h1} a_{h2} \ldots a_{hn}} \sum_{D_1=0}^{1} \cdots \sum_{D_n=0}^{1} (-1)^{D_1+\cdots+D_n} J_n\left(\sum_{j=1}^{n} a_{hj}\theta_{j,D_j} + b_h\right), \tag{32}$$

where we define $\theta_{j,0} = \theta_j$ and $\theta_{j,1} = \underline{\theta}_j$ $(j = 1, 2, \ldots, n)$, the upper and lower bounds of the integration intervals; the primitive $J_n(\cdot)$ is given by (31) in Theorem 1.

The marginal distribution functions $CDF_{\theta_j}(\theta_j)$ $(j = 1, \ldots, n)$ are now obtained by taking $\theta_l = \bar{\theta}_l \; \forall l = 1, \ldots, n; l \neq j$ in (32). The conditional density kernel of $(\theta_1, \ldots, \theta_j)$ given $(\theta_{j+1}, \ldots, \theta_n)$ is simply obtained by substituting the values $\theta_{j+1}, \ldots, \theta_n$ into (30); a kernel of the conditional CDF is given by (32) with $\sum_{l=j+1}^{n} a_{hl}\theta_l + b_h$ instead of b_h (and j instead of n).

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 functions; it seems that taking a few steps of the bisection method followed by the Newton–Raphson method works well in practice.

3.2.2. 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:

$$nn(\theta) = \exp\left(\sum_{h=1}^{H} c_h \, plin(a_h'\theta + b_h) + d\right), \tag{33}$$

where each element θ_j is restricted to a certain finite interval $[\underline{\theta}_j, \bar{\theta}_j]$ $(j = 1, \ldots, n)$. It is easy to perform Gibbs sampling from this distribution, as one can divide the domain of each θ_j $(j = 1, \ldots, n)$ 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_j is itself a piecewise-linear function of θ_j. 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 (22) are used here explicitly. A more detailed description of this procedure can be found in Hoogerheide et al. (2004).

Another possible method to draw from the Type 2 neural network density is auxiliary variable Gibbs sampling, which 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_j $(j = 1, \ldots, n)$. In the case of our Type 2 neural network the vector of latent variables u is $(H \times 1)$ where conditionally on θ the

[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.

u_h $(h = 1, \ldots, H)$ are independently drawn from uniform distributions:

$$u_h|\theta \sim U\left(0, \exp\left[c_h\, plin\left(\sum_{j=1}^{n} a_{hj}\theta_j + b_h\right)\right]\right), \quad h = 1, \ldots, H. \tag{34}$$

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

$$\theta_{j,LB}(u, \theta_{-j}) = \max\left\{\underline{\theta}_j, \max_{1 \leqslant h \leqslant H}\left\{\frac{1}{a_{hj}}\left(\frac{\log(u_h)}{c_h} - \frac{1}{2} - \left(\sum_{l=1, l\neq j}^{n} a_{hl}\theta_l + b_h\right)\right)\right| c_h a_{hj} > 0, 0 < \frac{\log(u_h)}{c_h} < 1\right\}\right\}, \tag{35}$$

$$\theta_{j,UB}(u, \theta_{-j}) = \min\left\{\bar{\theta}_j, \min_{1 \leqslant h \leqslant H}\left\{\frac{1}{a_{hj}}\left(\frac{\log(u_h)}{c_h} - \frac{1}{2} - \left(\sum_{l=1, l\neq j}^{n} a_{hl}\theta_l + b_h\right)\right)\right| c_h a_{hj} < 0, 0 < \frac{\log(u_h)}{c_h} < 1\right\}\right\}. \tag{36}$$

The derivations of these conditional distributions are given in Hoogerheide et al. (2004). Using auxiliary variable Gibbs sampling, 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.

3.2.3. Type 3 (mixture of t) neural network approximation

As we already remarked in the previous subsection, sampling from a Type 3 network, a mixture of t densities, only requires a drawing from the U(0, 1) distribution to determine which component is chosen, and a drawing from the chosen multivariate t distribution.

3.3. IS and the MH algorithm

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 IS and the MH algorithm, see footnote 2. We note that in the case of the Type 2 neural network the Gibbs sampler is used to obtain drawings; this case can be dealt with using a MH within Gibbs algorithm in which a MH step is considered after each time an element θ_j is drawn from its conditional neural network distribution.

4. Illustrative examples

In this section we consider the posterior distributions in IV regression models in order to compare the performance of the Type 3 (mixture of t densities) neural network sampling method (AdMit) with some other sampling methods.[9] First, consider the joint posterior of

[9] In the examples shown in this section the Type 1 and 2 networks performed worse than the Type 3 network. However, it is naive to expect one sampling method to dominate in all practical cases. A comparison of the

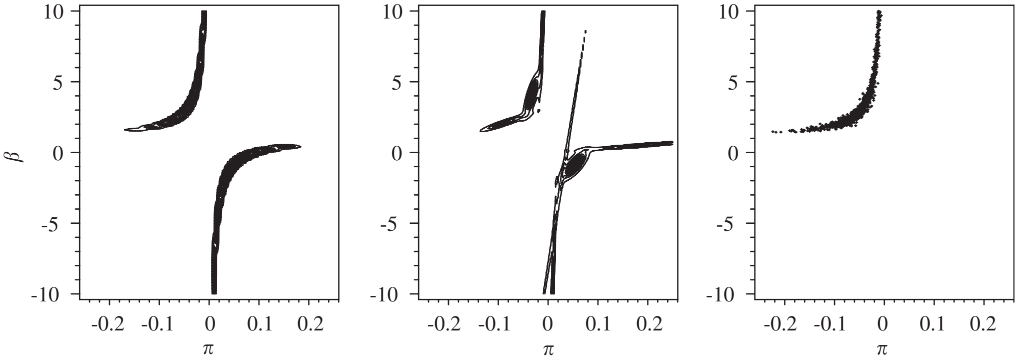

Fig. 6. Contour plots in the $\pi \times \beta$ plane: joint posterior of π and β in IV model for simulated data set with $\pi = 0.1$, $\rho = 0.99$ (left), and its Type 3 neural network approximation (middle); scatter plot of sample obtained by the Gibbs sampler (right).

π and β in (4) for the data set simulated from the model (1)–(2) with $\pi = 0.1$ (weak identification) and $\rho = 0.99$ (strong endogeneity), discussed in Section 2, truncated to the region

$$\{(\pi, \beta) | -0.25 \leqslant \pi \leqslant 0.25, -10 \leqslant \beta \leqslant 10\}. \tag{37}$$

The left panel of Fig. 6 shows its contour plot on this region (37). The contour plot of the Type 3 neural network approximation[10] is given by the middle panel of this figure; this contour plot 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 the right panel of Fig. 6. Of course, one can draw from the other ridge by choosing a different initial value, but it is not a trivial issue how to weight the results from the two ridges, i.e. it is not trivial to determine which part of the posterior probability mass is contained in each of both ridges.

Second, consider the joint posterior of $\pi = (\pi_1, \pi_2)'$ and β in (4) for $T = 50$ simulated data points from the model (1)–(2) with $\beta = 0$, $\sigma_{11} = \sigma_{22} = 1$, $\pi_1 = \pi_2 = 0.1$ (weak identification) and $\rho = 0.99$ (strong endogeneity), with $k = 2$ vectors of instruments consisting of i.i.d. N(0,1) drawings, truncated to the region

$$\{(\pi_1, \pi_2, \beta) | -0.5 \leqslant \pi_i \leqslant 0.5 \ (i = 1, 2), -10 \leqslant \beta \leqslant 10\}. \tag{38}$$

The middle panel of Fig. 7 shows the shape of an HPD credible set of (π_1, π_2, β) in the region (38) for this simulated data set. The left and right panels of Fig. 7 display the shapes of HPD credible sets in similar models with $T = 50$ simulated data points from the model (1)–(2) 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 two-dimensional distributions (ridges, bimodality and nearly elliptical shapes) also occur in these three-dimensional distributions.

(*footnote continued*)
performance of the neural network sampling methods will be reported in Hoogerheide and Van Dijk (2006). Some other examples illustrating the neural network sampling methods can be found in Hoogerheide et al. (2004).

[10] We constructed a mixture of 8 Student's *t* distributions with a sample of 50 000 IS weights with coefficient of variation of 2.1.

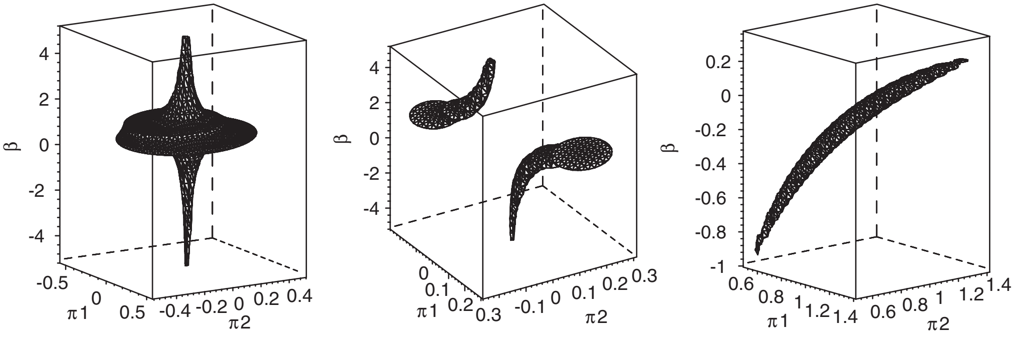

Fig. 7. Credible sets for parameters π_1, π_2, β in IV model (1)–(2) for simulated data sets from this model with strong endogeneity ($\rho = 0.99$) combined with either no ($\pi_1 = \pi_2 = 0$), weak ($\pi_1 = \pi_2 = 0.1$) or strong ($\pi_1 = \pi_2 = 1$) identification, respectively.

We use our AdMit procedure to construct a Type 3 neural network approximation, a mixture of 15 Student's t distributions, and use 1 000 000 drawings from it in IS and MH; see Table 2. The reported computing times correspond to an AMD Athlon$^{\text{TM}}$ 1.4 GHz processor. We have repeated the algorithms 20 times; Table 2 shows the standard deviations of the 20 estimates of $\text{E}(\pi_1)$, $\text{E}(\pi_2)$ and $\text{E}(\beta)$. The table also 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 $\text{E}(\pi_1)$, $\text{E}(\pi_2)$ and $\text{E}(\beta)$. The RNE is the ratio between (an estimate of) the variance of an estimator based on direct sampling and the IS estimator's estimated variance (with the same number of drawings). The RNE is an indicator of the efficiency of the chosen importance function; if target and importance density coincide the RNE equals one, whereas a very poor importance density will have an RNE close to zero.

The performance of AdMit-IS (in the same computing time) is compared with IS using a unimodal importance density, the Student's t distribution with $v = 1$ degree of freedom. In order to give the unimodal density a fair chance, the mode and scale are first iteratively updated four times as the estimated mean and covariance matrix of the target distribution in the previous step. The results are in Table 2. AdMit-IS gives standard deviations of the 20 estimates of $\text{E}(\pi_1)$, $\text{E}(\pi_2)$, $\text{E}(\beta)$ that are $2.3, 2.0, 2.2$ times as small, respectively, while the numerical standard errors are $1.9, 1.9, 3.4$ times as small for AdMit-IS. Also notice the huge differences between the RNEs (especially for the estimate of $\text{E}(\beta)$), the total weights of the 5% most influential points and the coefficients of variation of the weights in the two IS methods.

We compare the performance of AdMit-MH with the independence chain MH algorithm using a Student's t distribution with $v = 1$ degree of freedom, and with the random walk (RW) MH algorithm with candidate steps from a t_1 distribution. The scale (and mode) are first iteratively updated 4 times as the estimated covariance matrix (and mean) of the target distribution in the previous step. The results are in Table 2. AdMit-MH yields standard deviations of the 20 estimates of $\text{E}(\pi_1)$, $\text{E}(\pi_2)$, $\text{E}(\beta)$ that are $1.9, 1.9, 3.6$ times smaller than t_1 (independence chain) MH, and $1.6, 1.5, 1.3$ times smaller than RW MH. Also note that AdMit-MH has much higher acceptance rate and lower (first order) serial correlations in the MH chain.

Table 2
Sampling results for the non-elliptically shaped posterior distribution in the IV regression (1)–(2) with $k = 2$ instruments for simulated data with $\pi = (0.1, 0.1)'$ (weak identification), $\rho = 0.99$ (strong endogeneity)

	True values	AdMit IS	AdMit MH	Adaptive t_1 IS	Adaptive t_1 MH	Adaptive RW MH
$\mathrm{E}(\pi_1)$	0.0199	0.0200	0.0195	0.0203	0.0193	0.0206
(st.dev. 20×)		(1.2×10^{-4})	(1.9×10^{-4})	(2.8×10^{-4})	(3.7×10^{-4})	(2.9×10^{-4})
(num. std. error)		(1.6×10^{-4})		(3.1×10^{-4})		
[RNE]		[0.3622]		[0.0032]		
$\mathrm{E}(\pi_2)$	0.0157	0.0158	0.0153	0.0161	0.0152	0.0165
(st.dev. 20×)		(1.4×10^{-4})	(2.0×10^{-4})	(2.8×10^{-4})	(3.7×10^{-4})	(3.0×10^{-4})
(num. std. error)		(1.6×10^{-4})		(2.9×10^{-4})		
[RNE]		[0.3586]		[0.0034]		
$\mathrm{E}(\beta)$	0.6404	0.6357	0.6531	0.6327	0.6291	0.6121
(st.dev. 20×)		(0.0070)	(0.0110)	(0.0154)	(0.0394)	(0.0141)
(num. std. error)		(0.0065)		(0.0220)		
[RNE]		[0.2211]		[0.0006]		
$\sigma(\pi_1)$	0.0946	0.0945	0.0943	0.0946	0.0945	0.0946
$\sigma(\pi_2)$	0.0935	0.0934	0.0934	0.0935	0.0938	0.0935
$\sigma(\beta)$	3.0643	3.0745	3.0713	3.0682	3.0447	3.0816
Total time (s)		927	927	1067	1160	1138
Time construction NN (s)		598	598			
Time adapting scale (s)				88	106	83
Time sampling (s)		329	329	979	1054	1055
Drawings		1×10^6	1×10^6	30×10^6	30×10^6	50×10^6
Time/drawing (ms)		0.33	0.33	0.03	0.04	0.02
Coeff. var. IS weights		1.47		21.6		
5% largest IS weights (%)		27.3		99.999		
Acceptance rate MH (%)			32.5		0.4	2.3
Serial corr. π_1			0.66		0.995	0.994
Serial corr. π_2			0.66		0.995	0.994
Serial corr. β			0.72		0.996	0.996

Comparing the standard deviations of the estimates of $\mathrm{E}(\pi_1)$, $\mathrm{E}(\pi_2)$, $\mathrm{E}(\beta)$ in the five algorithms, AdMit-IS performs best: its standard deviations are about 1.5 times smaller than those of AdMit-MH, and at least twice as small as IS/MH with a t_1 importance/candidate density or the RW MH algorithm.

The Gibbs sampler failed in this example: the Gibbs sequence remained in one of the two ridges for 25 000 000 drawings (taking 1039 s).

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

Finally, consider the joint posterior of π and β in (4) for the data set simulated from the model (1)–(2) with $k = 1$ instrument with $\pi = 1$ (strong identification) and $\rho = 0$ (no endogeneity), discussed in Section 2, truncated to the region

$$\{(\pi, \beta) | -0.5 \leqslant \pi \leqslant 1.5, -10 \leqslant \beta \leqslant 10\}. \tag{39}$$

Fig. 1 shows its contour plot, which shows an elliptical shape. We construct a Type 3 neural network approximation, a mixture of two Student's t distributions. Many diagnostic

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

	True	AdMit	AdMit	Gibbs	RW MH	IS	MH	IS	MH
	values	IS	MH			t_1	t_1	normal	normal
E(π)	0.908	0.908	0.911	0.910	0.908	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(β)	−0.028	−0.025	−0.029	−0.029	−0.029	−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	0.093	0.088	0.087	0.087
$\sigma(\beta)$	0.106	0.105	0.102	0.104	0.105	0.105	0.105	0.102	0.102
corr(π, β)	0.017	0.041	−0.013	0.086	0.021	0.041	0.015	−0.019	−0.020
Total time (s)		20.8	20.9	0.03	0.64	0.03	0.11	0.11	0.12
Time construction NN (s)		20.7	20.7						
Time sampling (s)		0.05	0.16	0.03	0.64	0.03	0.11	0.11	0.12
Drawings		1000	2500	1000	40000	1000	2500	4000	4000
Time/drawing (ms)		0.05	0.06	0.03	0.02	0.03	0.04	0.03	0.03
Coeff. var. IS weights		0.797				0.797		0.163	
5% largest IS weights (%)		11.1				11.1		7.5	
Acceptance rate MH (%)			58.6		39.0		60.5		93.5
Serial corr. π			0.40	−0.02	0.85		0.38		0.11
Serial corr. β			0.39	−0.04	0.85		0.36		0.14

checks have been developed for assessing the convergence of the IS and MH methods; see e.g. Geweke (1989) for IS and Cowles and Carlin (1996) for MCMC methods. Here, a simple heuristic rule is used to obtain estimates of the means with (roughly) 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(π) and the difference between the two estimates of E(β) are both less than 0.005.[11] The results are in Table 3. We compare AdMit's performance with the Gibbs sampler, the RW MH algorithm with candidate steps from a t_1 distribution with scale matrix equal to minus the 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 target distribution the Gibbs sampler and 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 emphasizes that different sampling methods dominate in different cases; the neural network approach is especially useful for target densities with highly non-elliptical contours. Strategies to determine which method should be used in which situation are discussed in Hoogerheide and Van Dijk (2006).

[11] 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.

5. Conclusion

We have shown that the shape of Bayesian HPD credible sets is often non-elliptical in IV regression models with weak instruments and/or strong endogeneity. Structural inference is possible in the overidentified model but the credible sets may indicate large uncertainty. Unless one uses a truncated region of integration, 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 posterior probabilities and marginal densities in cases of distributions displaying such non-elliptical HPD 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 IS or the MH 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. In an illustrative example of 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 RW MH 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. One needs to develop a strategy in which a sophisticated network is specified for complex, non-elliptical densities, whereas 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 Hoogerheide and Van Dijk (2006).

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. An area of further research is to consider different flexible candidate density functions involving Hermite polynomials, see e.g. Gallant and Tauchen (1992) 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. 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 Bauwens et al. (2004) in a class of adaptive radial-based direction sampling methods (ARDS). A combination of ARDS 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.

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).

Finally, 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 et al. (1994).

Acknowledgements

We thank Andrew Chesher, two anonymous referees, and participants of seminars at Econometric Institute Rotterdam, Tokyo Metropolitan University, CORE, ESEM, JSM, CEF, and EC^2 meetings for comments on an earlier version of this paper. We are, in particular, indebted to Geert Dhaene for very helpful comments which led to substantial improvements. All remaining errors are the authors' responsibility.

References

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

Bauwens, L., Van Dijk, H.K., 1990. Bayesian limited information analysis revisited. In: Gabszewicz, J.J., et al. (Eds.), Economic Decision-Making: Games, Econometrics and Optimisation. North-Holland, Amsterdam, pp. 385–424.

Bauwens, L., Bos, C.S., Van Dijk, H.K., Van Oest, R.D., 2004. Adaptive radial-based direction sampling: some flexible and robust Monte Carlo integration methods. Journal of Econometrics 123, 201–225.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, second ed. Springer, New York.

Bos, C.S., Mahieu, R.J., Van Dijk, H.K., 2000. Daily exchange rate behaviour and hedging of currency risk. Journal of Applied Econometrics 15, 671–696.

Chao, J.C., Phillips, P.C.B., 1998. Bayesian posterior distributions in limited information analysis of the simultaneous equation model using Jeffreys' prior. Journal of Econometrics 87, 49–86.

Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91, 883–904.

Damien, P., Wakefield, J., Walker, S., 1999. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. Journal of the Royal Statistical Society B 61, 331–344.

Donaldson, R.G., Kamstra, M., 1997. An artificial neural network-GARCH model for international stock return volatility. Journal of Empirical Finance 4 (1), 17–46.

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

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

Edwards, R.G., Sokal, A.D., 1988. Generalization of the Fortuin–Kasteleyn–Swendsen–Wang representation and Monte Carlo algorithm. Physical Review D 38, 1012–2009.

Frühwirth-Schnatter, S., 2004. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. Econometrics Journal 7, 143–167.

Gallant, A.R., Tauchen, G., 1992. A nonparametric approach to nonlinear time series analysis: estimation and simulation. In: Brillinger, D., Caines, P., Geweke, J., Parzen, E., Rosenblatt, M., Taqqu, M.S. (Eds.), New Directions in Time Series Analysis Part II. Springer, New York, pp. 71–92.

Gallant, A.R., White, H., 1988. There exists a neural network that does not make avoidable mistakes. In: Proceedings of the Second Annual IEEE Conference on Neural Networks. IEEE Press, New York.

Gelman, A., Meng, X.-L., 1991. A note on bivariate distributions that are conditionally normal. The American Statistician 45, 125–126.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57, 1317–1339.

Hamilton, J.D., 1989. A new approach to the econometric analysis of nonstationary time series and business cycles. Econometrica 57, 357–384.

Hammersley, J., Handscomb, D., 1964. Monte Carlo Methods. Chapman & Hall, London.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Hecht-Nielsen, R., 1987. Kolmogorov mapping neural network existence theorem. In: Proceedings of the First Annual IEEE Conference on Neural Networks. IEEE Press, New York.

Hoogerheide, L.F., Van Dijk, H.K., 2006. Strategies to approximate posterior densities with neural networks: exploring a class of flexible sampling methods. Econometric Institute Report, Erasmus University Rotterdam, forthcoming.

Hoogerheide, L.F., Kaashoek, J.F., Van Dijk, H.K., 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.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Hutchinson, J., Lo, A., Poggio, T., 1994. A nonparametric approach to the pricing and hedging of derivative securities via learning networks. Journal of Finance 49, 851–889.

Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.

Kleibergen, F.R., Van Dijk, H.K., 1994a. Bayesian analysis of simultaneous equation models using noninformative priors. Discussion Paper TI 94-134, Tinbergen Institute, Rotterdam.

Kleibergen, F.R., Van Dijk, H.K., 1994b. On the shape of the likelihood/posterior in cointegration models. Econometric Theory 10 (3–4), 514–551.

Kleibergen, F.R., Van Dijk, H.K., 1998. Bayesian simultaneous equations analysis using reduced rank structures. Econometric Theory 14 (6), 701–743.

Kleibergen, F.R., Zivot, E., 2003. Bayesian and classical approaches to instrumental variable regression. Journal of Econometrics 114, 29–72.

Kloek, T., Van Dijk, H.K., 1978. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. Econometrica 46, 1–19.

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 28, 55–59 (Russian original in Doklady Akademii Nauk SSSR, 144, 953–956).

Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S., 1993. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6, 861–867.

Maddala, G.S., 1976. Weak priors and sharp posteriors in simultaneous equation models. Econometrica 44, 345–351.

Meng, X.-L., Wong, W.H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica 6, 831–860.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1091.

Paap, R., Van Dijk, H.K., 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.

Pesaran, M.H., Smith, R., 1995. Estimation of long-run relationships from dynamic heterogeneous panels. Journal of Econometrics 68, 79–113.

Shephard, N., 1996. Statistical aspects of ARCH and stochastic volatility. In: Cox, D.R., Hinkley, D.V., Barndorff-Nielson, O.E. (Eds.), Time Series Models with Econometric, Finance and Other Applications. Chapman & Hall, London.

Staiger, D., Stock, J.H., 1997. Instrumental variable regression with weak instruments. Econometrica 65, 557–586.

Strachan, R.W., Van Dijk, H.K., 2004. Valuing structure, model uncertainty and model averaging in VAR processes. Econometric Institute Report 2004-18, Erasmus University, Rotterdam.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Annals of Statistics 22, 1701–1762.

Van Dijk, H.K., 2003. On Bayesian structural inference in a simultaneous equation model. In: Stigum, B.P. (Ed.), Econometrics and the Philosophy of Economics. Princeton University Press, Princeton, NJ.

Van Dijk, H.K., Kloek, T., 1980. Further experience in Bayesian analysis using Monte Carlo integration. Journal of Econometrics 14, 307–328.

Van Dijk, H.K., Kloek, T., 1984. Experiments with some alternatives for simple importance sampling in Monte Carlo integration. In: Bernardo, J.M., Degroot, M., Lindley, D., Smith, A.F.M. (Eds.), Bayesian Statistics, Vol. 2. North-Holland, Amsterdam.

Zeevi, A.J., Meir, R., 1997. Density estimation through convex combinations of densities; approximation and estimation bounds. Neural Networks 10, 99–106.

Zellner, A., 1971. An Introduction to Bayesian Inference in Econometrics. Wiley, New York.

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