Hoogerheide, Lennart; van Dijk, Herman K.

Working Paper

Possibly Ill-behaved Posteriors in Econometric Models

Tinbergen Institute Discussion Paper, No. 08-036/4

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Hoogerheide, Lennart; van Dijk, Herman K. (2008) : Possibly Ill-behaved Posteriors in Econometric Models, Tinbergen Institute Discussion Paper, No. 08-036/4, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/86743

TI 2008-036/4
Tinbergen Institute Discussion Paper

Possibly Ill-behaved Posteriors in Econometric Models

Lennart Hoogerheide
Herman K. van Dijk

Econometric Institute, Erasmus University Rotterdam, and Tinbergen Institute.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

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

Possibly Ill-behaved Posteriors in Econometric Models: On the Connection between Model Structures, Non-elliptical Credible Sets and Neural Network Simulation Techniques *

Lennart Hoogerheide† & Herman K. van Dijk†

April 2008

Tinbergen Institute report 08-036/4

Abstract

Highly non-elliptical posterior distributions may occur in several econometric models, in particular, when the likelihood is allowed to dominate and information in the data is weak. This latter feature occurs frequently in empirical econometric analysis. Well-known cases are: instrumental variable models with weak instruments like the income-education models; vector autoregressive models with co-integration restrictions, widely used for the analysis of macroeconomic and financial time series; and mixture processes where one component is nearly non-identified like business cycle models with recessions and expansions as components of the mixture.

We explain the issue of highly non-elliptical posteriors in the context of a simple model for the effect of education on income using data from the well-known Angrist and Krueger (1991) study and discuss how a so-called Information Matrix or Jeffreys' prior may be used as a 'regularization prior' that in combination with the likelihood function yields posteriors with desirable properties. We also illustrate that the IV model and the vector autoregressive model with co-integration restrictions have a similar mathematical structure and thus this leads to similar posterior shapes.

In order to perform a Bayesian posterior analysis using simulation techniques in these models, one has to face the issue of finding a good candidate density

*Preliminary versions of this paper were presented at the 2007 ISI Conference in Lisbon, the 2008 MCMSki meeting in Bormio, and at the University of Montreal, Harvard University and Louisiana State. Helpful comments of several participants led to substantial improvements. The authors further thank David Ardia for useful suggestions. The second author gratefully acknowledges the hospitality of Harvard's Economics department where part of this paper was written and financial assistance from the Netherlands Organization of Research (grant 400-07-703).

†Econometric and Tinbergen Institutes, Erasmus University Rotterdam, The Netherlands.

for all classes of indirect sampling methods. In a recent paper – Hoogerheide, Kaashoek and Van Dijk (2007) – a class of neural network functions was introduced as candidate densities in case of non-elliptical posteriors.

In the present paper, the connection between canonical model structures, non-elliptical credible sets, and more sophisticated neural network simulation techniques is explored. As a preliminary step, three types of neural networks are applied to a bimodal distribution of Gelman and Meng (1991) and it is shown that the type of neural network that amounts to a mixture of Student's t densities clearly outperforms the two other types of networks in terms of computing time. Next, the performance of a mixture of Student's t distributions is compared with a Student's t distribution as a candidate for a 2-dimensional posterior distribution in a simple IV model for the effect of education on income, using data on men born in the state New York. Finally, an 8-dimensional bimodal posterior distribution is analyzed in a 2-regime mixture model for the real US GNP growth. In all examples considered in this paper, the mixture of Student's t distributions is clearly a much better candidate, yielding far more precise estimates of posterior means after the same amount of computing time, whereas the Student's t candidate almost completely misses substantial parts of the parameter space.

JEL classification: C11; C15; C45.
Keywords: instrumental variables; vector error correction model; Jeffreys' prior; mixture model; importance sampling; Markov chain Monte Carlo; neural network.

1 Introduction

There exist classes of statistical and econometric models where the joint and marginal posterior distributions of the parameters may have unknown analytical properties and non-elliptical Bayesian Highest Posterior Density [HPD] credible sets, see e.g. Berger (1985). Then it is not trivial to perform inference on the joint posterior distribution. This may have strong effects on the measurement of uncertainty of forecasts and of certain policy measures. The feature of non-elliptical posteriors occurs frequently in empirical econometric analysis. We mention here three cases. First, instrumental variable models with weak instruments like the income-education models which are relevant for government agencies responsible for compulsory schooling laws. Secondly, near unit root models and – more generally – vector autoregressive models with co-

integration restrictions, widely used for the analysis of macroeconomic and financial time series. For instance, in international financial markets, these models are used for hedging currency risk, and knowledge of a strongly non-elliptical credible set is important for the specification of an optimal hedging decision under risk. Thirdly, mixture processes where one component is nearly non-identified. As an example we consider business cycle models with recessions and expansions as components of the mixture. A detailed analysis of the literature is beyond the scope of the present paper. For some details on econometric models we refer to Imbens and Angrist (1994) and Bos, Mahieu and Van Dijk (2000) and the references cited there.

An important issue is that one may encounter great difficulties when trying to simulate (pseudo-) random draws from such a non-elliptical joint posterior distribution. Even if it is relatively easy to simulate random draws 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 investigate the ill-behaved posterior distributions that may occur in the IV regression model. We consider a simple, illustrative model for the measurement of the effect of education on income for two different data sets of Angrist and Krueger (1991). In this way, we also illustrate the effect of instrument strength on the posterior shapes, as the strength of the instrument differs considerably between the two data sets. We show the peculiar posterior shapes under the diffuse prior and explain the working of the Information Matrix or Jeffreys' prior as a 'regularization prior', that in combination with the likelihood function yields posteriors with desirable properties. Further, we illustrate that the similar mathematical structure of the instrumental variable model and the vector autoregressive model under cointegration restrictions leads to similar posterior shapes.

A second contribution of this paper is to extend the analysis of neural network sampling, introduced by Hoogerheide, Kaashoek and Van Dijk (2007) [henceforth HKVD]. These methods allow for sampling from a target (posterior) distribution that may be multi-modal or skew. In other words, this 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 function is constructed that approximates the target density (kernel). In the second step this neural network function is embedded in a Metropolis-Hastings [MH] or importance sampling [IS] algorithm.[1] With respect to

[1]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

the first step we emphasize that an important advantage of neural network functions is their 'universal approximation property'. That is, under certain conditions neural network functions can provide approximations of any square integrable function to any desired accuracy.[2] 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 MH or IS, the candidate density function or 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 draws 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.

In this paper, we extend the HKVD analysis as follows. First, we apply three types of neural networks to a bimodal, conditionally normal distribution of Gelman and Meng (1991) in order to compare the computing times required for the three neural network sampling methods. We analyze why the neural network that amounts to a mixture of Student's t densities outperforms the two other types of networks in terms of computing time, and explain how this candidate density - that approximates the posterior distribution - is iteratively constructed. Second, we compare the mixture of Student's t distributions with a unimodal t distribution as a candidate distribution for a 2-dimensional posterior distribution in a simple IV model for the effect of education on income, using data on men born in the state New York. Third, we compare the mixture of t distributions with a t distribution as a candidate distribution for an 8-dimensional posterior distribution in a 2-regime mixture model for the real US GNP growth.

The outline of the paper is as follows. In Section 2 we consider the model structure and the shapes of posterior densities in a simple IV regression model, and similar posterior shapes in the VECM. In Section 3, we consider the three types of neural network functions that can be used as candidate densities in case of non-elliptical posteriors. We explain why some of the well-known possible drawbacks of neural networks do not play a role in this application. Section 4 provides a comparison of the performance of the three neural network functions as candidate densities for a bimodal, conditionally normal distribution of Gelman and Meng (1991). In Section 5, we compare the mix-

Van Dijk (1978) and is further developed by Van Dijk and Kloek (1980, 1984) and Geweke (1989).

[2]Kolmogorov (1957) and Hecht-Nielsen (1987) establish general theoretical capabilities. Proofs concerning neural network approximations for specific configurations can be found in *e.g.* Gallant and White (1988), Hornik, Stinchcombe and White (1989) and Leshno, Lin, Pinkus and Schocken (1993).

ture of Student's t distributions with a Student's t distribution as a candidate for a posterior in a simple IV model. We illustrate that it is worthwhile to 'invest' some computing time in an accurate candidate density or importance function, as this investment may become very 'profitable' in the sense of much quicker convergence or more reliable sampling results. In Section 6, the sampling performance of the mixture of t distributions is analyzed as a candidate distribution for an 8-dimensional posterior distribution in a 2-regime mixture model for the real US GNP growth. The proposed method in Section 6 differs from the approach in Section 5 that heavily relies on the evaluation of Hessian matrices, which can be troublesome in higher dimensions or in situations with pronounced boundaries in the parameter space. The proposed algorithm is also different from the method of HKVD, in the sense that it 'learns' the neural network candidate density in a somewhat more intelligent manner. The results for an 8-dimensional highly non-elliptical posterior suggest the method's useful applicability in higher dimensions. Finally, we show the shapes of the likelihood function in a particular mixture model, illustrating that the prior of e.g. Frühwirth-Schnatter (2001) can also be interpreted as a 'regularization' prior that eliminates the likelihood function's 'spikes'. Section 7 gives concluding remarks and some topics for further research on which we intend to report in the near future.

2 The issue of ill-behaved posterior densities in the instrumental variables (IV) regression model, illustrated for the measurement of the effect of education on income

A well-known example of the use of instrumental variables in econometrics is the measurement of the effect of education on income, the (monetary) return on education. Measuring the effect of education on income, is a matter of great importance for several decision processes. For example, the results of such analysis are relevant for government agencies responsible for compulsory schooling laws, for school districts considering changes in school entrance policies and also for parents deciding when to enroll their children to school. However, a problem is that intellectual capabilities, which are usually not observed, not only influence education but also directly affect income. Therefore, a simple regression of income on the number of years of education may lead to incorrect conclusions. For example, more intelligent students find school

less difficult and may choose to obtain more schooling to signal their high ability. So, even if extra years of education have no effect on income, people with higher education will on average have higher incomes because of their higher abilities. Therefore, one may expect that an ordinary regression of income on the years of education leads to an upward bias, i.e. an overestimated effect of education on income. Further, the (often unobserved) intellectual capabilities, income and education level of the parents may also cause an upward bias, as the parents' characteristics may also influence the education level and have a direct effect on income. For example, it may be the case that children of more intelligent and higher educated parents on average learn more at home. Another problem is the measurement error in reported education. First, usually only the completed (integer) number of years of education is reported. Second, people may misreport their education spell.[3] If the measurement error would be the only problem, one would expect that a simple regression of income on education would result in a downward bias, i.e. an underestimated effect of education on income, as the part of the variation in education that is merely due to measurement error does not lead to variation in income.

A method for solving these problems is the use of instrumental variables. These instrumental variables must be correlated with education but uncorrelated with latent capabilities (and measurement errors). Intuitively, in this way one focuses on the direct effect of education on income, while other effects on income are filtered out. However, it is hard to find variables that are correlated with education but uncorrelated with intellectual capabilities. Angrist and Krueger (1991) use American data and suggest using quarter of birth to form instrumental variables. These instruments exploit that students born in different quarters have different average education spells. This results since most school districts require students to have turned age six by a certain date, a so-called 'birthday cutoff' which is typically near the end of the year, in the year they enter school, whereas compulsory schooling laws compel students to remain at school until their sixteenth, seventeenth or eighteenth birthday. This asymmetry between school-entry requirements and compulsory schooling laws compels students born in certain months to attend school longer than students born in other months: students born earlier in the year enter school at an older age and reach the legal dropout age after less education. Hence, for students who leave school as soon as the schooling laws allow for it, those born in the first quarter have on average attended school for three quarters less than those born in the fourth quarter.

[3]Siegel and Hodge (1968) find that the correlation between individuals' education reported in two surveys is only 0.933.

Angrist and Krueger (1991) use three data sets on men born in three decades, emphasizing results for the data set on 329509 men born in the years 1930-1939. For the latter data set we consider a simple, illustrative model for persons $i = 1, \ldots, N$:

$$y_i = x_i \beta + \varepsilon_i \tag{1}$$

$$x_i = z_i \Pi + v_i \tag{2}$$

with y_i the log weekly wage in 1979, x_i the number of years of education, and $z_i = 1$ if person i is born in quarter 2, 3 or 4, and $z_i = 0$ if person i is born in quarter 1. The variables y_i, x_i, z_i are taken in deviation from their means, so that no constant terms occur in (1) and (2). The parameter β is the average effect of one extra year of education on income: on average, one more year of schooling results in an increase of income of approximately 100β %. The (scalar) parameter Π is the difference in the mean education spell between men born in quarter 2, 3 or 4 and men born in quarter 1. The error terms ε_i and v_i are assumed to be independent across observations and jointly normally distributed: $(\varepsilon_i, v_i)' \sim N(0, \Sigma)$.

We consider both the case with the whole data set and the case in which we only use data on 29015 men born in the state New York. Especially in the latter case, the quarter-of-birth instrument is very weak. As an indication, for the New York data the first stage F-statistic is 0.55 (with p-value 0.46), whereas for the whole US data set this is 67.57 (with p-value 0.00). Figure 1 shows the data.

First, we consider the following diffuse prior

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

which is used by Zellner (1971) and Drèze (1976) for particular values of h.

Given the model (1)-(3), one can easily derive the likelihood function and the posterior density kernel of (β, Π, Σ). Choosing $h = 3$ in the prior density kernel (3) and 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, leads to the following joint posterior kernel of (β, Π):

$$p(\beta, \Pi | y, x, Z) \propto \left| \begin{array}{cc} (y - x\beta)'(y - x\beta) & (y - x\beta)'(x - Z\Pi) \\ (x - Z\Pi)'(y - x\beta) & (x - Z\Pi)'(x - Z\Pi) \end{array} \right|^{-N/2}, \tag{4}$$

where y and x are $N \times 1$ vectors, Z is an $N \times k$ matrix with k the number of instruments, and Π is a $k \times 1$ vector; in our simple example we have $k = 1$. The marginal posterior of β, derived by Drèze (1976, 1977), see also Bauwens and Van Dijk (1990), is given

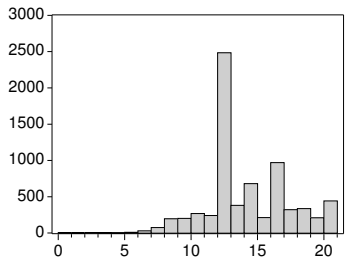
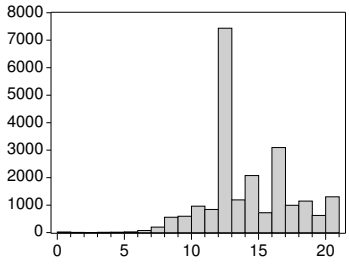
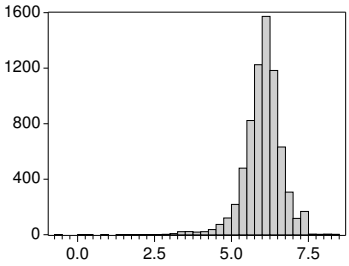
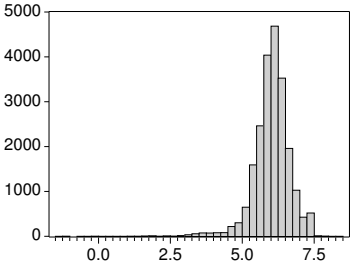
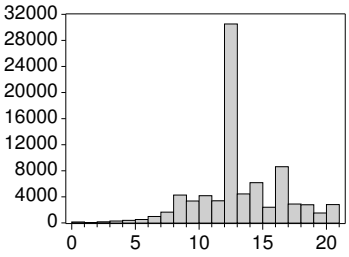
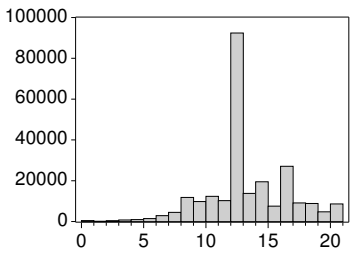
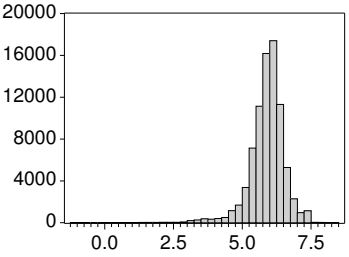
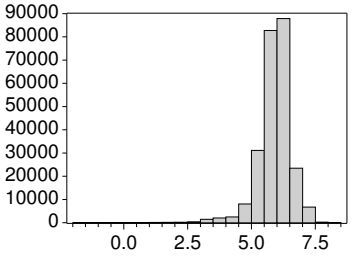
	born in 1st quarter	born in quarter 2-4
Men born in the state New York: ($N = 29015$)		
years of education		
	mean = 13.67, st.dev. = 3.18	mean = 13.71, st.dev. = 3.15
log income		
	mean = 6.016, st.dev. = 0.647	mean = 6.013, st.dev. = 0.664
Men born in the US: ($N = 329509$)		
years of education		
	mean = 12.69, st.dev. = 3.31	mean = 12.80, st.dev. = 3.27
log income		
	mean = 5.892, st.dev. = 0.681	mean = 5.903, st.dev. = 0.678

Figure 1: Data on education and income for samples of men born in 1930-1939, which were also used by Angrist and Krueger (1991). (Obviously, the New York data are a subset of the US data.) The differences between mean education and income for different quarters of birth are slightly larger for the US data.

by:

$$p(\beta|y,x,Z) \;\propto\; \frac{[(y-x\beta)'(y-x\beta)]^{-(N-1)/2}}{[(y-x\beta)'M_Z(y-x\beta)]^{-(N-k-1)/2}} \tag{5}$$

with $M_Z = I - Z(Z'Z)^{-1}Z'$. Kleibergen and Van Dijk (1994, 1998) derived the marginal posterior of Π as:

$$\begin{aligned} p(\Pi|y,x,Z) \;\propto\; & [(x-Z\Pi)'(x-Z\Pi)]^{-(N-1)/2}\,(\Pi'Z'M_xZ\Pi)^{-1/2}\times \\ & \times\left(\frac{\Pi'Z'M_{[y\,x]}Z\Pi}{\Pi'Z'M_xZ\Pi}\right)^{-(N-1)/2} \end{aligned} \tag{6}$$

These posterior distributions have several peculiar properties:

(a) **Local non-identification at $\Pi = 0$:** The marginal posterior of Π has an asymptote at $\Pi = 0$ because of the term $(\Pi'Z'M_xZ\Pi)^{-1/2}$. In the case of $k = 1$ instrument, the posterior is not integrable over neighborhoods around $\Pi = 0$. (See Kleibergen and Van Dijk (1994, 1998).)

(b) **Regular posterior behavior of β when irrelevant instruments are added:** The marginal posterior of β becomes tighter if (possibly irrelevant) instruments are added. Moments exist up to the order of overidentification $(k-1)$; for $k = 1$, the marginal posterior of β is improper. (This result appeared in an informal way in Maddala (1976), commenting on Drèze (1976).)

These pathologies stem from the local non-identification of β when $\Pi = 0$, which is most easily seen from the restricted reduced form corresponding to the structural form (1)-(2):

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} = \begin{pmatrix} \beta \\ 1 \end{pmatrix} \Pi'\, z_i + \begin{pmatrix} v_{1i} \\ v_i \end{pmatrix} \tag{7}$$

with $v_{1i} = v_i\beta + \varepsilon$ and $(v_{1i}, v_i)' \sim N(0,\Omega)$. Figure 2 illustrates these pathologies for the data of New York and the whole US. For the joint posterior kernel of β and Π for New York data, a substantial 'ridge' is visible at $\Pi = 0$; the marginal posterior of Π is completely dominated by the asymptote at $\Pi = 0$. On the other hand, for the US data, the shapes are nearly elliptical, which reflects that in this case the quarter-of-birth instrument is less weak. The peak around the posterior mode[4] is high compared with the ridge around $\Pi = 0$, so that the latter is not visible. Still, the joint posterior has a non-integrable ridge at $\Pi = 0$, as can be seen from the asymptote at $\Pi = 0$ for the marginal posterior of Π.

[4]In this simple example, the posterior mode is given by $(\beta,\Pi) = (\hat{\beta}_{2SLS}, \hat{\Pi}_{OLS}) = (y'z/x'z, x'z/z'z)$.

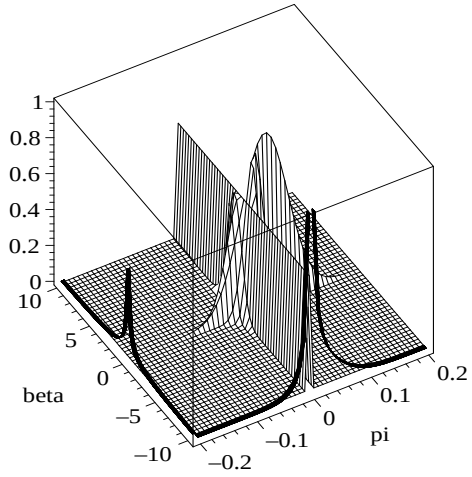
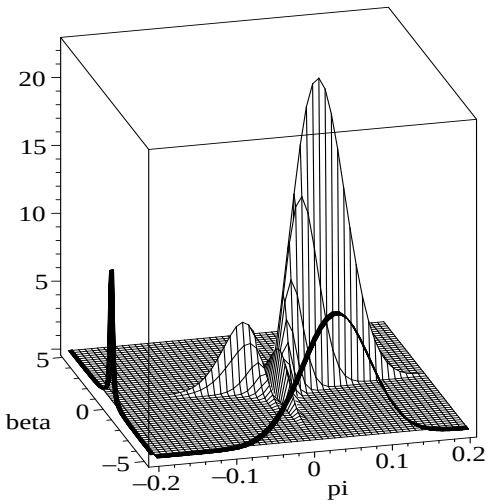
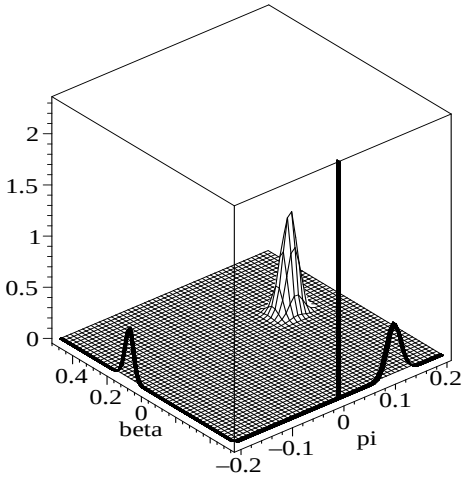
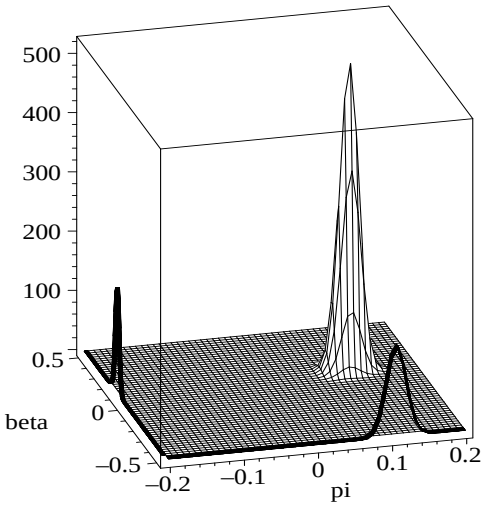
	Posterior density kernel $p(\beta, \Pi \mid data)$ under diffuse prior	Posterior density $p(\beta, \Pi \mid data)$ under Jeffreys' prior
data of New York ($N = 29015$)		
data of USA ($N = 329509$)		

Figure 2: Posterior density kernels for the simple IV model (1)-(2) for measurement of the effect of education on income (β) using the difference in mean education between men born in quarters 2-4 and quarter 1 (Π). The graphs show the joint posterior kernel of β, Π. At the axes, the marginal posterior kernels of β and Π are shown.

We now consider the Information Matrix or Jeffreys prior. The Jeffreys prior, the square root of the determinant of the information matrix, is given by:

$$p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-2} \, (\Pi' Z' Z \Pi)^{1/2} \, \sigma_{22.1}^{-(k-1)/2} \tag{8}$$

with $\sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$, for the structural form (1)-(2), or equivalently by:

$$p(\beta, \Pi, \Omega) \propto |\Omega|^{-2} \, (\Pi' Z' Z \Pi)^{1/2} \, ((\beta \; 1) \Omega^{-1} (\beta \; 1)')^{(k-1)/2} \tag{9}$$

for the corresponding restricted reduced form (7); see Appendix A of Hoogerheide, Kleibergen and Van Dijk (2007) for a derivation of this Jeffreys prior.

The factor $(\Pi' Z' Z \Pi)^{1/2}$ is 0 for $\Pi = 0$, which reflects that in the restricted reduced form β only occurs in the product $\Pi\beta$, so that for $\Pi = 0$ the model contains no information on β. Hence for $\Pi = 0$ the likelihood is constant over values of β, so that the first and second order derivatives of the log-likelihood with respect to β are zero, and the determinant of the information matrix, minus the expectation of the Hessian of the log-likelihood, is 0 for zero values of Π.

Intuitively speaking, the factor $(\Pi' Z' Z \Pi)^{1/2}$ in the prior 'cancels' the asymptote of the posterior at $\Pi = 0$ so the posteriors are proper even in case of a just identified model.

The $((\beta \; 1) \Omega^{-1} (\beta \; 1)')^{(k-1)/2}$ factor in the prior influences the tail behavior of the marginal posterior of β and makes it independent of the number of instruments k such that it has Cauchy type tails.

Note that for $k = 1$ instrument the Jeffreys prior (8) reduces to

$$p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-2} |\Pi|, \tag{10}$$

which is simply the diffuse prior in (3) with $h = 4$ multiplied with $|\Pi|$. One interpretation of this Information Matrix or Jeffreys prior is that a priori one prefers a strong instrument; that is, Π is preferred to be large (in absolute sense). An intuitively appealing explanation is that this Jeffreys prior is just a *'regularization prior'* that does not immediately reflect prior beliefs, but in combination with the likelihood function yields posteriors with desirable properties (in the sense that the aforementioned peculiar properties resulting from the diffuse prior do not occur).

Notice that also for $k > 2$ the factor $(\Pi' Z' Z \Pi)^{1/2}$ in the prior takes high values for (in absolute sense) large elements of Π, while in this case the $((\beta \; 1) \Omega^{-1} (\beta \; 1)')^{(k-1)/2}$ factor takes high values for (in absolute sense) large values of β. In the likelihood of the (restricted reduced form of) the IV model, it is the occurrence of the product $\Pi\beta$

that causes points (Π, β) with Π and β both attaining extremely large values to have small posterior probability.

Figure 2 illustrates the posterior shapes under the Jeffreys prior for the data of New York and the US. For the US data, the graphs look similar to the graphs under the diffuse prior, except for the disappearance of the asymptote at $\Pi = 0$ for the marginal posterior of Π. For the New York data, the differences with the posterior shapes under the diffuse prior are huge. Under the Jeffreys prior, there is no ridge or asymptote at $\Pi = 0$, and the tails of the marginal posterior of β are thinner (and integrable). Also notice that, although the *joint* posterior kernel of β, Π tends to 0 for $\Pi \to 0$, the *marginal* posterior of Π does not drop in neighborhoods of $\Pi = 0$: for $\Pi \to 0$ the lower values of the posterior density kernel $p(\beta, \Pi|y, x, Z)$ are compensated by the fact that for $\Pi \to 0$ the posterior $p(\beta, \Pi|y, x, Z)$ becomes less sensitive with respect to changes in β, as β only occurs in the likelihood in the product $\Pi\beta$. In other words, the marginal posterior probability mass of Π does not decrease for $\Pi \to 0$, this posterior probability mass is just spread over a wider range of values for β. Finally, note that although the Jeffreys prior 'cures' some of the peculiar properties under the diffuse prior, the posterior may still display non-elliptical shapes such as bimodality.

It should be noted that the model above is much simpler than the models considered by Angrist and Krueger (1991); for example, Angrist and Krueger (1991) also include dummies for the direct effect of state and year of birth on education and income. Using a model of Angrist and Krueger (1991), Hoogerheide and Van Dijk (2006) show that the results for US data depend to a large extent on the data of three states (Arkansas, Kentucky, Tennessee). For many states, including the state of New York, the quarter of birth instrument has hardly any value. We note that there exists an extensive literature on the interpretation of IV estimands as local average treatment effects [LATE]. For more details, we refer to Angrist, Imbens and Rubin (1996) and Imbens and Angrist (1997a, 1997b).

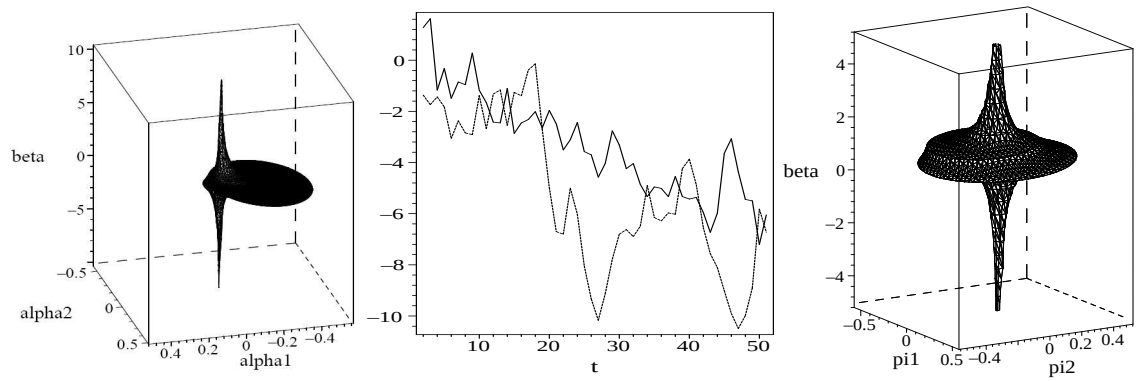


Figure 3: A Highest Posterior Density credible set for the parameters α_1, α_2, $\tilde{\beta}$ in the VECM under a diffuse prior for simulated data from a VECM with $\alpha_1 = -0.05$, $\alpha_2 = 0.05$, $\tilde{\beta} = 1$ (left); the simulated data from the VECM (middle); an HPD credible set in an IV model in a similar simulation experiment (right)

Similarity of mathematical structure and posterior shapes in IV model and Vector Error Correction Model

Consider the following restricted reduced form of an IV model with 2 instruments z_{1i}, z_{2i} $(i = 1, \ldots, N)$, and a simple vector error correction model (VECM) under a cointegration restriction for 2 variables y_{1t}, y_{2t} $(t = 1, \ldots, T)$:

$$\text{IV: } \begin{pmatrix} y_i \\ x_i \end{pmatrix} = \overbrace{\begin{pmatrix} \beta \\ 1 \end{pmatrix} (\pi_1 \;\; \pi_2)}^{\text{reduced rank}} \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} v_{1i} \\ v_i \end{pmatrix}$$

$$\text{VECM: } \begin{pmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{pmatrix} = \overbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} (1 \;\; -\tilde{\beta})} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

which have in common that they contain a parameter matrix with reduced rank. In both models, local non-identification plays a role. In the IV model, the parameter β is not identified for $\pi_1 = \pi_2 = 0$, whereas in the VECM the parameter $\tilde{\beta}$ is not identified for $\alpha_1 = \alpha_2 = 0$.

We now consider a simulation experiment with $\alpha_1 = -0.05$, $\alpha_2 = 0.05$, $\tilde{\beta} = 1$, so that there is slow adjustment towards the cointegration relation $y_1 = y_2$, $(\varepsilon_{1t}, \varepsilon_{2t}) \sim N(0, I)$, for a rather small data set $(T = 50)$. The left panel of Figure 3 shows a Highest Posterior Density (HPD) credible set for $(\alpha_1, \alpha_2, \beta)$ under a diffuse prior similar to the diffuse prior for the IV model, for $-0.5 < \alpha_j < 0.5$ $(j = 1, 2)$, $-10 < \tilde{\beta} < 10$. The middle panel of Figure 3 shows the simulated data from the VECM. The right panel

shows approximately the same non-elliptical posterior shapes for a similar simulation experiment in the IV model.

3 Neural network sampling methods

In the previous section, it was shown that the posterior distributions in the IV model and VECM may be highly non-elliptical. This property is shared by many other models, such as the class of mixture models, which will be considered in the sequel of this paper. A problem in the presence of highly non-elliptical posterior shapes is that if one desires to investigate properties of the posterior density $p(\theta|data)$ (of a m-dimensional parameter vector θ), using indirect sampling methods as Importance Sampling (IS) or the independence chain Metropolis-Hastings (MH) algorithm, then using an elliptical candidate distribution gives slow convergence and/or incorrect results.

In such a situation, one possible approach is to use a neural network function as the candidate density. The three types of neural network functions introduced by HKVD are as follows.

The first specification, the *Type 1* neural network, is a three-layer feed-forward neural network, a *multi-layer perceptron* [MLP], with arctangent activation function:

$$nn(\theta) = \sum_{h=1}^{H} c_h \arctan\left(\sum_{k=1}^{m} a_{hk}\theta_k + b_h\right) + d \tag{11}$$

where H reflects the number of hidden cells of the network, and a_{hi}, b_h, c_h, d (with $h = 1, \ldots, H$, $k = 1, \ldots, m$) represent the network weights that have to be estimated. Figure 4 shows (for the case with $m = 2$, $H = 2$) the network diagram representing the Type 1 neural network.

The reason for choosing the arctangent function is that it can be analytically integrated infinitely many times. This property makes the neural network, in the role of a density kernel on a bounded region, easy to sample from, because each marginal and conditional cumulative distribution function [CDF] can be analytically derived. For details, we refer to HKVD and Hoogerheide (2006).

HKVD suggest the following procedure to 'learn' the weights of the Type 1 neural network approximation to a certain target posterior density kernel $p(\theta|data)$. First, obtain a set of draws θ^j $(j = 1, \ldots, n)$ from the uniform distribution on the bounded region to which we restrict the random variable $\theta \in \mathbb{R}^m$ to take its values. Then approximate the target density kernel $p(\theta|data)$ with a neural network by minimizing

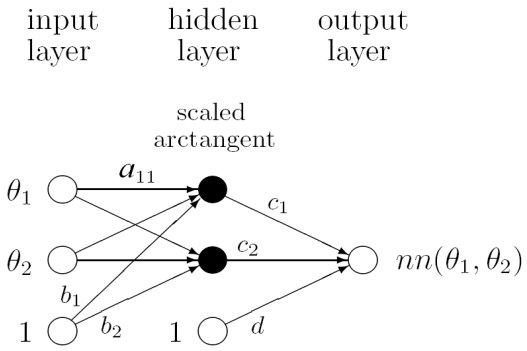


Figure 4: Network diagram corresponding to the Type 1 neural network, a multi-layer perceptron with arctangent activation function, in case of $m = 2$ inputs and $H = 2$ hidden cells.

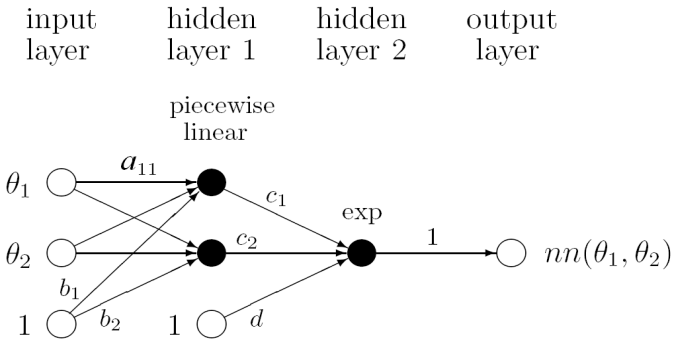


Figure 5: Network diagram corresponding to the Type 2 neural network, which applies the exponential transformation to the output of a multi-layer perceptron with piecewise-linear activation function, in case of $m = 2$ inputs and $H = 2$ hidden cells.

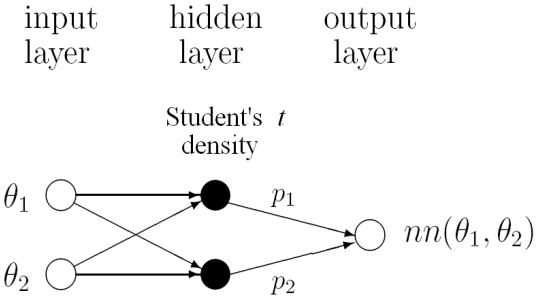


Figure 6: Network diagram corresponding to the Type 3 neural network, which amounts to a mixture of Student's t densities, as a 3-layer Radial Basis Function (RBF) network, in case of $m = 2$ inputs and $H = 2$ mixture components.

the sum of squared residuals:

$$SSR(A,b,c,d) = \sum_{j=1}^{n} \left(p(\theta^j|data) - nn\big(\theta^j \big|\, A,b,c,d\big)\right)^2, \tag{12}$$

We choose the most parsimonious neural network, i.e. the one with the smallest number H of 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 $p(\theta|data)$ and $nn(\theta)$. In the case of a Type 1 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 (12), for example $-M \sum_{j=1}^{n} I\{nn(\theta^j) < 0\}\ nn(\theta^j)$ where M is a constant large enough to make nn non-negative in all points θ^j $(j = 1, \ldots, n)$. 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 accuracy of the approximation and the non-negativity. For example, one can check the squared correlation R^2 between $nn(\theta)$ and $p(\theta|data)$ for a much larger set of points than the 'estimation set', and one can look for the (global) minimum of $nn(\theta)$ by running a minimization procedure starting with several initial values.

The second specification, the *Type 2* neural network, is a network of which the output is the exponential function of a three-layer feed-forward neural network function with piecewise-linear activation function:

$$nn(\theta) = \exp\left[\sum_{h=1}^{H} c_h \ \text{plin}\left(\sum_{k=1}^{m} a_{hk}\theta_k + b_h\right) + d\right] \tag{13}$$

with

$$\text{plin}(x) = \begin{cases} 0 & x < -1/2 \\ x + 1/2 & -1/2 \le x \le 1/2 \\ 1 & x > 1/2 \end{cases}, \qquad x \in \mathbb{R}. \tag{14}$$

Figure 5 shows (for the case with $m = 2$, $H = 2$) the network diagram representing the Type 2 neural network. The idea behind this specification is that the candidate density kernel (13) allows for easy Gibbs sampling (see Geman and Geman (1984)); (13) can be analytically integrated with respect to a θ_k $(k = 1, \ldots, m)$, after which one uses analytical inversion of the conditional CDF to generate the next draw in the Gibbs sequence. Since the draws from the *Type 2* network are obtained as a Gibbs sequence, the corresponding Metropolis-Hastings algorithm is a so-called 'Metropolis-Hastings within Gibbs' method.

Again, the network weights can be 'learned' by minimizing (12); for the Type 2 network no penalty function is required, as the exponential function implies that non-negativity is automatically taken care of.

The third specification, the *Type 3* neural network, is a mixture of Student's t densities:

$$nn(\theta) = \sum_{h=1}^{H} p_h \; t(\theta|\mu_h, \Sigma_h, \nu), \tag{15}$$

where p_h $(h = 1, \ldots, H)$ are the probabilities (satisfying $p_h \geq 0$, $\sum_{h=1}^{H} p_h = 1$) of the Student's t components and where $t(\theta|\mu_h, \Sigma_h, \nu)$ is an m-variate Student's t density with mode vector μ_h, scaling matrix Σ_h, and ν degrees of freedom:

$$t(\theta|\mu_h, \Sigma_h, \nu) = \frac{\Gamma((\nu+m)/2)}{\Gamma(\nu/2)(\pi\nu)^{m/2}} \, |\Sigma_h|^{-1/2} \left(1 + \frac{(\theta-\mu_h)'\Sigma_h^{-1}(\theta-\mu_h)}{\nu}\right)^{-(\nu+m)/2}. \tag{16}$$

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. The Type 3 network can be interpreted as a radial basis function (RBF) network; Figure 6 shows (for the case with $m = 2$, $H = 2$) the corresponding network diagram.

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

First, compute the mode μ_1 and scale Σ_1 of the first Student's t distribution in the mixture as $\mu_1 = \text{argmax}_\theta \; p(\theta|data)$, the mode of the target distribution, and Σ_1 as minus the inverse Hessian of $\log p(\theta|data)$ evaluated at its mode μ_1. Then draw a set of points θ^j $(j = 1, \ldots, n)$ from the 'first stage neural network' $nn(\theta) = t(\theta|\mu_1, \Sigma_1, \nu)$, with small ν to allow for fat tails.[5] After that add components to the mixture, iteratively, by performing the following steps:

Step 1: Compute the importance sampling weights $w(\theta^j) = p(\theta^j|data)/nn(\theta^j)$ $(j = 1, \ldots, n)$. 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 $w(\theta^j)$ $(j = 1, \ldots, n)$.

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

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 $nn(\theta)$ will be used as the candidate distribution.[6] Otherwise, go to step 2.

Step 2: Add another Student's t distribution with density $t(\theta|\mu_h, \Sigma_h, \nu)$ to the mixture with $\mu_h = \text{argmax}_\theta\ w(\theta) = \text{argmax}_\theta\{p(\theta|data)/nn(\theta)\}$ and Σ_h equal to minus the inverse Hessian of $\log w(\theta) = \log p(\theta|data) - \log nn(\theta)$ evaluated at 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^j with the highest weight $w(\theta^j)$ in the sample $\{\theta^j | j = 1, \ldots, n\}$. The idea behind this choice of μ_h and Σ_h is that the new Student's 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.

Step 3: 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, draw n points θ_h^j from each component $t(\theta|\mu_h, \Sigma_h, \nu)$ $(h = 1, \ldots, H)$. Then minimize $E[w(\theta)^2]/E[w(\theta)]^2$, where:

$$E[w(\theta)^k] = \frac{1}{n}\sum_{j=1}^{n}\sum_{h=1}^{H} p_h\ w\big(\theta_h^j\big)^k \quad (k = 1, 2), \qquad w\big(\theta_h^j\big) = \frac{p(\theta_h^j|data)}{\sum_{l=1}^{H} p_l\ t\big(\theta_h^j|\mu_l, \Sigma_l, \nu\big)}. \tag{17}$$

Step 4: Draw a sample of n points θ^j $(j = 1, \ldots, n)$ from our new mixture of Student's t distributions, $nn(\theta) = \sum_{h=1}^{H} p_h\ t(\theta|\mu_h, \Sigma_h, \nu)$, and go to step 1; in order to draw a point from the density $nn(\theta)$ first use a draw 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 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

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

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 Student's 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 parameter vector $\theta \in \mathbb{R}^m$ takes its values.

If the region of integration of the parameters θ is bounded, it may occur in step 2 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 is chosen as the point θ^j with the highest weight $w(\theta^j)$ in the sample $\{\theta^j | j = 1, \ldots, n\}$, Σ_h is obtained as the matrix of estimated second moments around μ_h for a certain 'residual distribution' with density kernel:

$$res(\theta) = \max\{p(\theta|data) - \tilde{c}\, nn(\theta), 0\}, \tag{18}$$

where $\tilde{c}$ is a constant.[7] We take $\max\{., 0\}$ to make it a (non-negative) density kernel. This Σ_h is easily obtained by importance sampling with the current $nn(\theta)$ as the candidate density, using the sample θ^j $(j = 1, \ldots, n)$ from $nn(\theta)$ that we already have. In the case of a bounded region of integration, HKVD suggest obtaining μ_h and Σ_h as the mean and covariance matrix of the 'residual distribution' with density kernel (18). However, this may result in a μ_h in a region with already enough candidate probability mass, which does not occur when choosing μ_h as the point θ^j with the highest weight $w(\theta^j)$.

During the past 20 years many results on the approximation capabilities of neural networks have been published. For example, Hornik, Stinchcombe and White (1989) show that 3-layer feed-forward networks with an arbitrary sigmoid activation function can approximate any square integrable function (given sufficiently many hidden cells). This implies that the Type 1 and Type 2 networks can yield accurate approximations to a wide variety of density (kernel) functions. Further, 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

[7] There are two issues relevant for the choice of $\tilde{c}$. First, the new Student's t density should appear exactly at places where $nn(\theta)$ is too small (relative to $p(\theta|data)$), i.e. the scale Σ_h should not be too large. Second, there should be enough points θ^j with $w(\theta^j) > \tilde{c}$ in order to make Σ_h nonsingular.

densities in (15) falls within their framework.

Finally, note that two of the well-known possible drawbacks of neural networks, the 'black box' property and the danger of 'overfitting', are no disadvantages for this application. First, the aforementioned types of neural networks are obviously 'black boxes' in the sense that the working is not immediately clear, as the values of the individual network weights have no straightforward interpretation. However, only a reasonable approximation of the target posterior is desired, no interpretation of the network weights is required. Second, in our application there is no danger of 'overfitting', where not only a structural process is captured, but also random noise is 'fitted'. For the 'data' used in the learning process consist of (posterior density kernel) function evaluations *without random noise*.

4 A comparison of the performance of different neural network functions as candidate densities: conditionally normal distribution of Gelman and Meng (1991)

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 sections we have $\theta = (X_1, X_2)'$.

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)):

$$p(x_1, x_2) \propto \exp\left(-\frac{1}{2}\left[Ax_1^2x_2^2 + x_1^2 + x_2^2 - 2Bx_1x_2 - 2C_1x_1 - 2C_2x_2\right]\right), \tag{19}$$

where A, B, C_1 and C_2 are constants. Equation (19) can be rewritten as:

$$p(x_1, x_2) \propto \exp\left(-\frac{1}{2}\left[Ax_1^2x_2^2 + (x-\mu)'\Sigma^{-1}(x-\mu)\right]\right), \tag{20}$$

with:

$$\mu = \left[\frac{BC_2 + C_1}{1-B^2}, \frac{BC_1 + C_2}{1-B^2}\right]' \qquad \Sigma^{-1} = \begin{pmatrix} 1 & -B \\ -B & 1 \end{pmatrix},$$

so the term $Ax_1^2x_2^2$ causes deviations from the bivariate normal distribution. We consider the symmetric case in which $A = 1$, $B = 0$, $C_1 = C_2 = 3$, with conditional distributions

$$X_1|X_2 = x_2 \sim N\left(\frac{3}{1+x_2^2}, \frac{1}{1+x_2^2}\right) \qquad X_2|X_1 = x_1 \sim N\left(\frac{3}{1+x_1^2}, \frac{1}{1+x_1^2}\right). \tag{21}$$

For the Type 1 and 2 networks, we restrict the variables X_1 and X_2 to the interval [-2.5,7.5]. This restriction does not affect our estimates, as the probability mass outside this region is negligible.

The contour plots of the neural network approximations[8] are given by Figure 7,

[8]We constructed a Type 1 network with $H = 50$ hidden neurons, $R^2 = 0.9966$ 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; the $H = 13$ hidden neurons result from deleting the (almost) irrelevant hidden neurons from a network of $H = 25$ neurons. We also constructed a mixture of $H = 4$ Student's t distributions with a sample of 1000 IS weights with coefficient of variation equal to 0.87 (and in which the 5% most influential points have 11.6% weight).

together with the contour plot of the target density. These contour plots confirm that the three classes of neural networks are able to provide reasonable approximations to the target density. Figure 7 clearly suggests that the Type 1 (MLP) neural network provides the best approximation. Especially compared with the Type 3 (mixture of t) network, its approximation is clearly more accurate. However, a substantial drawback is the computing time required for the construction of the Type 1 approximation: this takes over 120 seconds (on an Intel Centrino$^{\text{TM}}$ Duo Core processor), whereas the 'learning' of the Type 3 network only takes less than 1 second. The construction of the Type 1 network takes relatively much time, as relatively many hidden cells ($H = 50$) are required to provide a reasonable Type 1 neural network approximation. Figure 8 illustrates how the AdMit procedure iteratively constructs an approximating candidate density, a mixture of four t densities, in four steps.

Given the constructed neural network approximations, we sample from these networks and use the samples in IS or the (independence chain) MH algorithm. Many diagnostic checks have been developed for assessing the convergence of the IS or MH method; see *e.g.* Kloek and Van Dijk (1978) and Geweke (1989) for the IS method and Cowles and Carlin (1996) and Brooks and Roberts (1998) for MCMC methods. Several diagnostic checks for investigating the convergence of IS and MCMC methods are also discussed by Hoogerheide, Van Dijk and Van Oest (2008). In this example, 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.[9] The results are in Table 1. 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 (an estimate of) the variance of an estimator based on direct sampling and the IS estimator's estimated variance (with the same number of draws). The RNE is an indicator of the efficiency of the chosen importance function; if target and impor-

[9]The number of draws 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 draws.

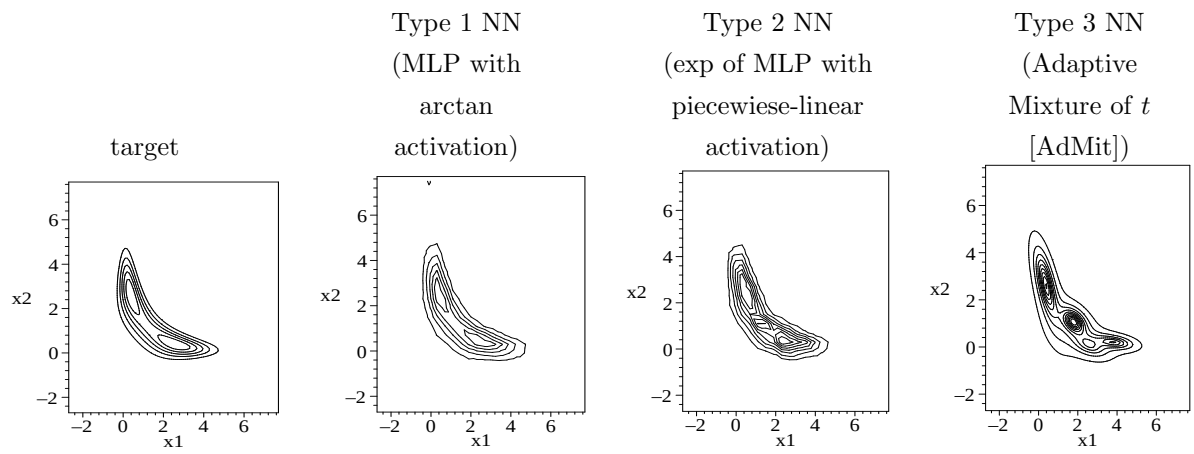


Figure 7: Contour plots: the target distribution, a conditionally normal bivariate distribution of Gelman and Meng (1991) in (21) (first), and its Type 1 (second), Type 2 (third), and Type 3 (fourth) neural network approximation.

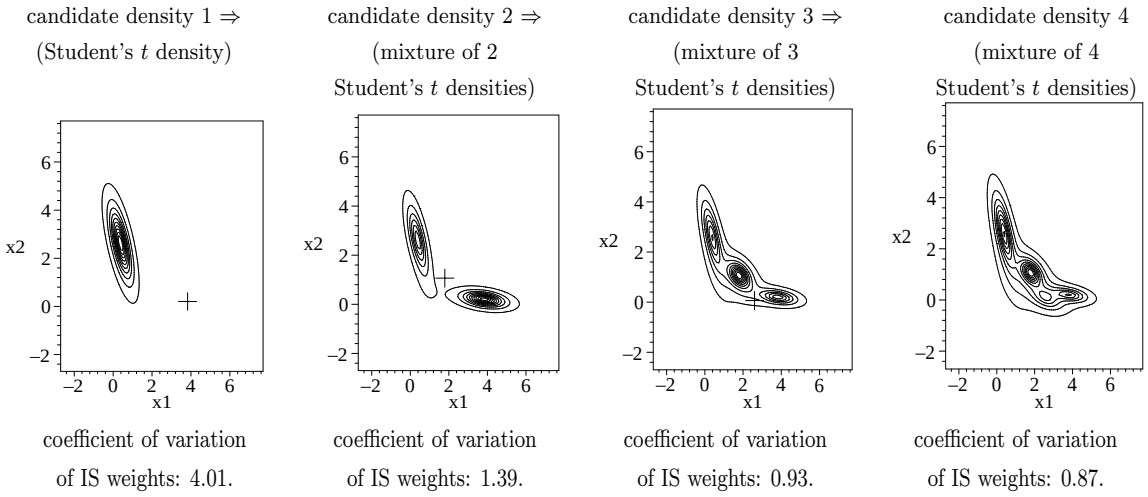


Figure 8: Illustration of the Adaptive Mixture of t [AdMit] procedure for constructing a Type 3 (mixture of t) neural network approximation to a target density, a bimodal conditionally normal distribution of Gelman and Meng (1991) in (21). In this case, a candidate density is constructed in four steps. The cross denotes the point at which the importance 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, which is the mode of the next Student's t distribution in the candidate mixture distribution. Below each panel the coefficient of variation, the standard deviation divided by the mean, of the importance sampling weights is reported.

Table 1: Neural network based sampling results for the conditionally normal bivariate distribution of Gelman and Meng (1991) in (21), which is depicted in the first panel of Figure 7. The IS and MH methods based on the Type 3 neural network, the mixture of Student's t densities, require much less computing time than the methods using Type 1 and 2 networks.

		importance function / candidate density					
		Type 1 NN (MLP with arctan activation)		Type 2 NN (exp of MLP with piecewise-linear activation)		Type 3 NN (Adaptive Mixture of t [AdMit])	
	real values	IS	MH	IS	MH	IS	MH
$E(X_1)$	1.459	1.487	1.504	1.472	1.433	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	1.459	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	1.236	1.245
$\sigma(X_2)$	1.234	1.239	1.235	1.223	1.244	1.242	1.235
$\rho(X_1, X_2)$	-0.760	-0.764	-0.766	-0.755	-0.757	-0.759	-0.759
total time		142.8 s	142.8 s	36.9 s	44.4 s	0.7 s	0.7 s
time construction NN		125.1 s	125.1 s	34.8 s	34.8 s	0.6 s	0.6 s
time sampling		17.7 s	17.7 s	2.1 s	9.6 s	0.1 s	0.1 s
draws		5000	5000	10000	40000	10000	10000
time/draw		3.5 ms	3.5 ms	0.21 ms	0.24 ms	0.01 ms	0.01 ms
5% IS weights		6.3 %		7.2 %		12.9 %	
coeff. var. IS weights		0.382		0.239		0.840	
acc. rate			84.6%		90.0 %		52.7 %
serial corr. X_1			0.15	0.65	0.73		0.45
serial corr. X_2			0.14	0.67	0.72		0.45

tance density coincide the RNE equals one, whereas a very poor importance density will have an RNE close to zero.[10]

The total weight of the 5% most influential points is below 15% for the three IS algorithms and the values of the RNE are rather high, confirming the quality of the importance density. The rather high MH acceptance rates above 50% reflect the quality of the neural network as a candidate density in the MH algorithm.

If we look at the computing times required for generating the samples, we conclude that AdMit-IS and AdMit-MH (based on the Type 3 network) are the winners in this

[10]The numerical standard error and RNE of Geweke (1989) are not reported for the Type 2 network, as the candidate draws are not independent, because these are generated by Gibbs sampling.

example. Not only does the 'learning' of the network take much less time than for the other networks, also sampling is performed far more quickly. Especially, sampling from a Type 1 network is rather slow as this requires a numerical method, such as the Newton-Raphson method, in order to invert the CDF. Further note that, whereas the approximation of the Type 2 network is somewhat better than that of the Type 3 network, more MH draws are required when using a Type 2 candidate. The reason for this is the higher serial correlation between the draws in this 'MH within Gibbs' approach. We conclude that this example clearly indicates the superiority of the Type 3 (mixture of t) network over the other two types: the slightly lower quality of the candidate as an approximation to the target density is easily compensated by the higher speed of both the 'learning' and the sampling.

The methods using Type 1 and Type 2 networks, especially the IS procedure for the Type 2 network, may become competitive if (much) better optimization techniques are used. Several different optimization methods than the used back-propagation method have been discussed in literature. For example, White (1989) shows that a particular back-propagation implementation is not efficient and discusses a two-step procedure that has better convergence properties.

The Type 1 network has the interesting property that the integral of its functional form can be evaluated analytically. Next to that, the moments can be derived analytically, see appendix 2.A.3 of Hoogerheide (2006). This means that if one can construct a Type 1 neural network that provides an (almost) perfect fit to the target density, then one can analytically evaluate the moments of the target distribution without the use of any Monte Carlo integration procedure. However, using a simple back-propagation technique, it is extremely time consuming to find a network with almost perfect fit. Application of optimization techniques that are specifically designed for neural network learning to the Type 1 and Type 2 network is a topic for further research.

5 A comparison of the performance of a mixture of Student's t densities with other candidate densities: a highly non-elliptical posterior in a simple IV model

The purpose of this section is to compare the performance of the mixture of Student's t densities, the Type 3 network, as a candidate density with a simple Student's t distribution in the presence of a highly non-elliptical posterior. As the target distribution we consider the posterior of the parameters Π, β in the simple IV model (1)-(2) under the diffuse prior for $N = 29015$ data on men born in New York in 1930-1939. The posterior density kernel is shown in the top-left panel of Figure 2 and the left panel of Figure 9. The reason for this choice of the target distribution in this example is simply that it has highly non-elliptical shapes because of the 'ridge' around $\Pi = 0$. We restrict the domain of Π, β to finite intervals, $(\Pi, \beta) \in [-0.2, 0.2] \times [-10, 10]$, as otherwise this posterior distribution is improper. The posterior density is well approximated by a mixture of 6 Student's t densities; see Figure 9. The first steps of the AdMit method are depicted in Figure 10.

The first columns of Table 2 give sampling results for the AdMit candidate density, the mixture of 6 Student's t densities. Further, Table 2 gives sampling results for a Student's t candidate density with mode and scale adapted to the posterior distribution in a preliminary run. The final two columns give results for the Student's t candidate density around the posterior mode (with scale matrix equal to minus the inverted Hessian of the log-posterior evaluated at the mode). Notice that the numbers of draws are chosen in such a way that the total amount of computing time is approximately the same among the three methods.[11]

Note that the IS and MH methods with the AdMit candidate yield estimates of the posterior means with a higher precision: for both the estimates of $E[\beta]$ and $E[\Pi]$ the numerical standard error is more than two times smaller than under the Student's t candidate distributions. Especially, under the Student's t distribution around the posterior mode the numerical standard error for the estimate of $E[\beta]$ is much worse.

[11] The numerical standard errors for the Metropolis-Hastings algorithm are estimated by the method of Andrews (1991), using a quadratic spectral (QS) kernel and pre-whitening as suggested by Andrews and Monahan (1992). The corresponding relative numerical efficiency (RNE) is the inverse of the inefficiency factor (IF), the MH estimator's (estimated) variance divided by the variance under direct sampling (using the same number of draws).

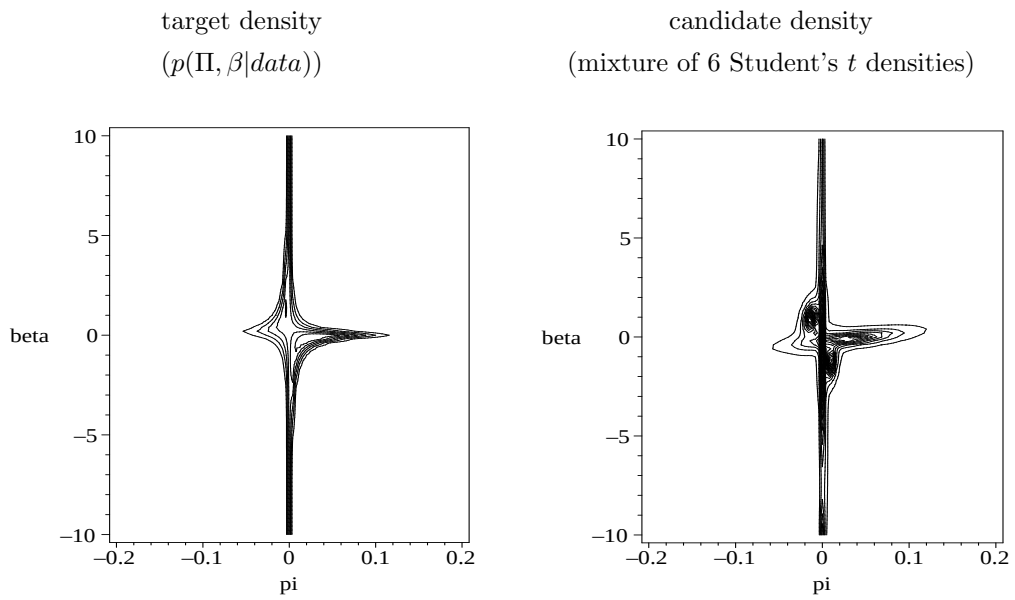


Figure 9: Contour plots: Posterior density kernel for parameters Π, β *in simple IV model (1)-(2) under the diffuse prior (3) for measurement of the effect of education on income (*β*), using as an instrument the difference in mean education between men born in quarters 2-4 and quarter 1* (Π)*, using* $N = 29015$ *data on men born in the state New York in 1930-1939 (left). Approximating candidate density, a mixture of 6 Student's t densities (right).*

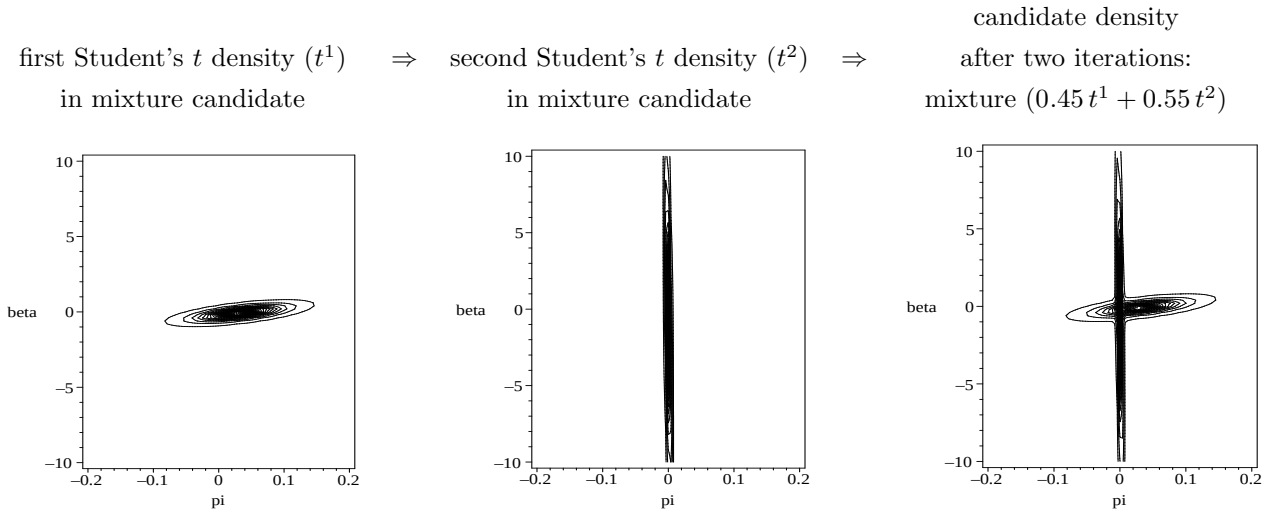


Figure 10: First steps of Adaptive mixture of t [AdMit] method in IV model (1)-(2) for data on men born in the state New York in 1930-1939 (under diffuse prior): first Student's t distribution of the mixture (around posterior mode) (left), second Student's t distribution of the mixture (middle), mixture of first and second Student's t distributions (right)

Table 2: Sampling results for different candidate densities: estimated posterior moments for simple IV model (1)-(2) under the diffuse prior (3) for measurement of the effect of education on income (β), using as an instrument the difference in mean education between men born in quarters 2-4 and quarter 1 (Π), using $N = 29015$ data on men born in the state New York in 1930-1939. (The (target) posterior density is depicted in the top-left panel of Figure 2 and the left panel of Figure 9). The IS and MH methods based on the Type 3 network, the mixture of Student's t densities, yield much more precise results in approximately the same computing time.

	importance function / candidate density					
	Type 3 NN (Adaptive Mixture of t [AdMit])		Student's t with adapted mode and scale		Student's t around posterior mode	
	IS	MH	IS	MH	IS	MH
$E(\beta)$	-0.0086	-0.0174	0.0001	-0.0012	0.2647	0.1023
(num. std. error)	(0.0052)	(0.0066)	(0.0112)	(0.0139)	(0.1563)	(0.2225)
[RNE]	[0.3866]	[0.2397]	[0.0277]	[0.0183]	[0.0002]	$[7.6 \cdot 10^{-5}]$
$E(\Pi)$	0.0080	0.0080	0.0080	0.0081	0.0079	0.0078
(num. std. error)	$(3.6 \cdot 10^{-5})$	$(5.5 \cdot 10^{-5})$	(0.0001)	(0.0001)	(0.0002)	(0.0002)
[RNE]	[0.4519]	[0.1920]	[0.0605]	[0.0305]	[0.0074]	[0.0047]
$\sigma(\beta)$	3.2265	3.2314	3.2266	3.2547	3.4033	3.3674
$\sigma(\Pi)$	0.0241	0.0241	0.0242	0.0242	0.0240	0.0237
total time	34.2 s		37.4 s		36.4 s	
time construction NN	18.1 s					
time adapting mode, scale			1.0 s			
time sampling	16.1 s		36.4 s		36.4 s	
draws	1 mln		3 mln		3 mln	
time/draw	0.016 ms		0.012 ms		0.012 ms	
coeff. var. IS weights	1.09		3.01		22.56	
5% largest IS weights	19.0 %		46.6 %		66.9 %	
acceptance rate MH		42.08 %		16.90 %		16.70 %
serial corr. β		0.627		0.937		0.994
serial corr. π		0.577		0.898		0.708

It should be noted that more than half of the computing time required for the results of IS or the MH algorithm using the AdMit candidate, is needed for 'learning' the candidate distribution, the mixture of 6 Student's t distributions. However, the RNE's are much higher for the AdMit candidate density, so that 1 million of AdMit draws are much more valuable than 3 millions of draws from the Student's t candidate distribution. The idea of the construction of a good candidate as an 'investment' is illustrated in Figure 11. Until 18.1 seconds the AdMit method is only constructing a candidate, while after 1 second (required for adapting the mode and scale to the target density) the IS approach with a Student's t candidate is already sampling. However, once the AdMit-IS method starts sampling, it soon outperforms IS with a Student's t candidate: the lines cross at 19.9 seconds, at a precision of $1/\text{var}(\widehat{E(\beta)}) = 4191.4$ (at a standard deviation of $\text{st.dev}(\widehat{E(\beta)}) = 0.0154$). AdMit-IS only requires 1.8 seconds to catch up with the 18.9 seconds of sampling of IS with a t candidate; the 'increase of precision per second of sampling' is about 10 times larger for AdMit-IS. The *increase of precision per second of sampling* for the IS estimator of the posterior mean of θ_k, the k-th element of θ, is given by:

$$\frac{\partial[1/\text{var}(\widehat{E(\theta_k)})]}{\partial \,\text{time}} = \frac{\#\text{draws per s}}{\text{var}(\theta_k)} \,\text{RNE}_{E(\theta_k)} \tag{22}$$

where the RNE (relative numerical efficiency) is the ratio between the (estimated) precision of the IS estimator of $E(\theta_k)$ and (an estimate of) the precision of an estimator of $E(\theta_k)$ based on direct sampling (with the same number of draws), see Geweke (1989). The *increase of precision per second of sampling* for the posterior mean of β is therefore given by

$$\frac{\partial[1/\text{var}(\widehat{E(\beta)})]}{\partial \,\text{time}} = \frac{1/(0.016 \cdot 10^{-3})}{(3.2265)^2}\, 0.3866 = 2321.0$$

for AdMit-IS;

$$\frac{\partial[1/\text{var}(\widehat{E(\beta)})]}{\partial \,\text{time}} = \frac{1/(0.012 \cdot 10^{-3})}{(3.2265)^2}\, 0.0277 = 221.7$$

for IS using a Student's t distribution with adapted mode and scale;

$$\frac{\partial[1/\text{var}(\widehat{E(\beta)})]}{\partial \,\text{time}} = \frac{1/(0.012 \cdot 10^{-3})}{(3.2265)^2}\, 0.0002 = 1.6$$

for IS using a Student's t distribution around the posterior mode.[12] So, if one desires

[12]Note that we use the same, most precise (AdMit-IS) estimate of $\text{st.dev}(\beta)$ of 3.2265 in the formulas. Moreover, notice that the time per draw is only a factor of approximately 4/3 larger for the mixture of 6 Student's t distributions than for the t distribution, because the evaluation of (mixtures of) t densities takes relatively little time. In these IS approaches, most time is required for evaluating the target density kernel.

to obtain an estimator of the posterior mean of β with standard deviation st.dev($\widehat{E(\beta)}$) smaller than 0.0154, then AdMit-IS is the better choice as in this case AdMit-IS needs (possibly much) less computing time. On the other hand, if one only needs a less precise estimator of the posterior mean of β, then IS with a Student's t candidate may be a better choice. However, one should bear in mind that the latter only holds, if the Student's t distribution covers the whole region containing posterior probability mass with enough candidate probability mass. For the probability that important regions of the parameter space, such as distant modes in case of a multi-modal posterior, are 'missed', is much smaller if one uses the AdMit procedure. So, next to the convergence speed of the sampling results, an advantage of the AdMit approach is the higher robustness, i.e. a higher reliability that the whole posterior distribution is covered by the candidate.

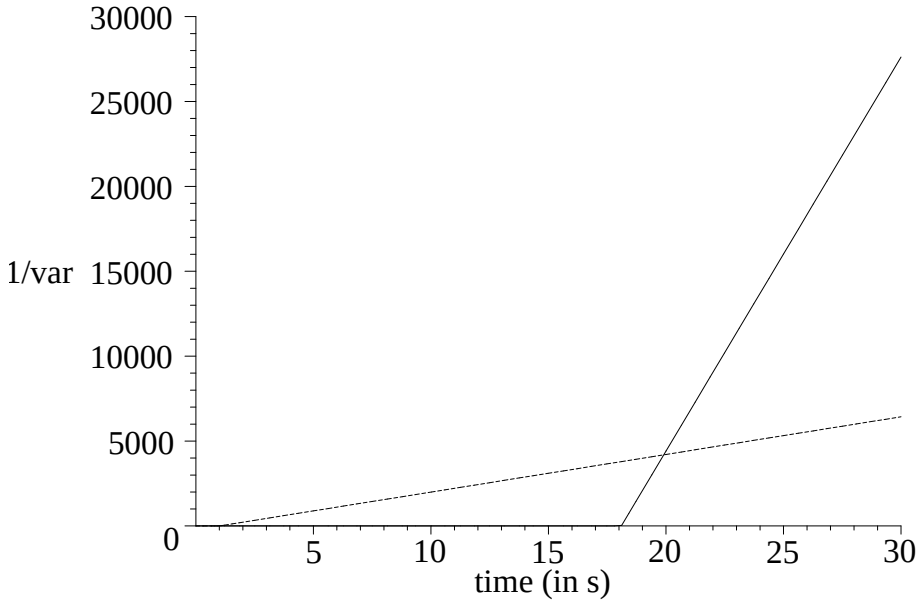


Figure 11: IV model (1)-(2) for data on men born in the state New York in 1930-1939 (under diffuse prior): precision (1/variance) of the IS estimator of the posterior mean of β for different candidate distributions, as a function of the computing time: Student's t candidate density with scale and mode adapted to target density (dashed line); AdMit (mixture of t) candidate density (solid line), which requires 18.1 seconds to be 'learned' but after that needs less than 2 seconds to outperform the Student's t candidate density.

6 Posterior shapes in a 2-regime mixture model for real US GNP growth: comparing candidate distributions for an 8-dimensional posterior

In this section we consider the posterior shapes in a 2-regime mixture model for real US GNP growth. We use this example model in order to compare candidate distributions in case of a highly non-elliptical, 8-dimensional posterior in a parameter space with a restricted domain. This indicates how poor a unimodal (Student's t) candidate distribution may perform in such situations, and how much quicker convergence of sampling results can be obtained by using a mixture of Student's t candidate distribution.

The used method differs from the approach in Section 5 that heavily relies on the evaluation of Hessian matrices, which can be troublesome in higher dimensions or in situations with pronounced boundaries in the parameter space, where the latter is the case in this example. The results for the 8-dimensional highly non-elliptical posterior suggest the method's useful applicability in higher dimensions.

We note that in this empirical example the mixture process refers to the data space. However, such mixture processes may give rise to bimodal or skew posterior distributions, i.e. non-elliptical shapes in the parameter space. In this example, we consider a mixture model with two AR(2) regimes for real US GNP growth:

$$\begin{aligned} y_t &= \begin{cases} \beta_{11} + \beta_{12} y_{t-1} + \beta_{13} y_{t-2} + \varepsilon_t & \text{with probability } p, \\ \beta_{21} + \beta_{22} y_{t-1} + \beta_{23} y_{t-2} + \varepsilon_t & \text{with probability } 1-p, \end{cases} \\ \varepsilon_t &\sim N(0, \sigma^2), \end{aligned} \tag{23}$$

where y_t denotes the (annualized) quarterly growth rate. The data consist of $T = 231$ observations from the first quarter of 1950 to the third quarter of 2007; see Figure 12. We emphasize that model (23) is used for illustrative purposes only. Investigating possible misspecification of (23) due to the Great Moderation in volatility observed since the early nineteen-eighties is beyond the scope of the present paper.

Note that we have an 8-dimensional vector $\theta = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23}, \sigma, p)'$. The prior for p is $U(0,1)$, while the prior for σ is taken proportional to $1/\sigma$, which amounts to specifying a uniform prior for $\log(\sigma)$. The priors for β_{i1} $(i = 1, 2)$ are chosen uniform on the interval $[-4, 4]$; for β_{i2}, β_{i3} $(i = 1, 2)$ the prior is chosen uniform on the interval $[-1, 1]$.[13] For identification, it is imposed that $\beta_{11} < \beta_{21}$.

[13] In order to obtain a proper posterior distribution for $\beta_{i1}, \beta_{i2}, \beta_{i3}$ $(i = 1, 2)$, we need to specify a proper prior for these parameters. Intuitively speaking, the reason is that there is a probability of

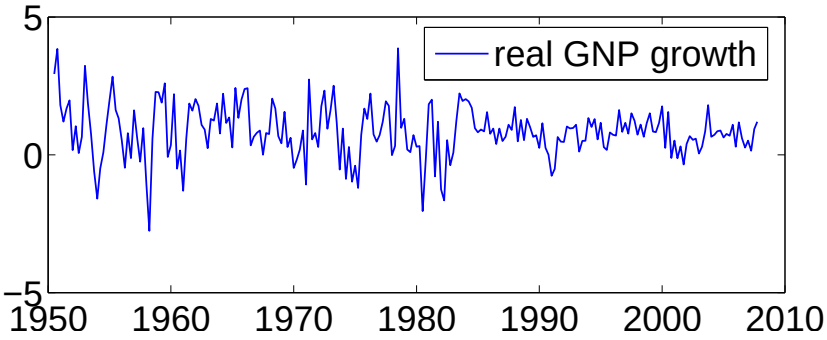


Figure 12: US real Gross National Product: (annualized) growth rates in percents. The data are seasonally adjusted. Source: US Department of Commerce, Bureau of Economic Analysis.

The constraint $\beta_{11} < \beta_{21}$ implies that in step 2 of the AdMit approach the importance weight function may often attain its supremum at this boundary (or at a boundary for $\beta_{i1} = \pm 4$ $(i = 1, 2)$, $\beta_{i2} = \pm 1$, or $\beta_{i3} = \pm 1$ $(i = 1, 2)$). So, the Hessian of the logarithm of the weight function, evaluated at its supremum, may often not be positive definite. Therefore, μ_h $(h = 2, 3, \ldots)$ is chosen as the point θ^j with the highest weight $w(\theta^j)$ in the current sample $\{\theta^j | j = 1, \ldots, n\}$, and Σ_h is obtained as the matrix of estimated second moments around μ_h for the 'residual distribution' with kernel in (18).

The first columns of Table 3 give sampling results for the AdMit candidate density, a mixture of 8 Student's t densities. Further, Table 3 gives sampling results for a Student's t candidate density with mode and scale that have been adapted to the posterior distribution in a preliminary run. The final two columns give results for the Student's t candidate density around the posterior mode (with scale matrix equal to minus the inverted Hessian of the log-posterior, evaluated at the mode). Notice that the numbers of draws are chosen in such a way that the total amount of computing time is approximately the same among the three methods. Note that the IS and MH methods with the AdMit candidate yield estimates of the posterior means with a higher precision: for all estimated posterior means the numerical standard error is smaller than under a Student's t candidate distribution. Under the AdMit-IS approach all numerical standard errors are more than 3 times smaller than under the other IS methods. For 4 parameters the AdMit-IS and AdMit-MH methods deliver numerical standard errors that are over 10 times smaller than under a Student's t candidate.[14]

$(1 - p)^T$ (a probability of p^T) that none of the observations belong to the first (second) regime, in which case the posterior of the corresponding parameters is simply given by the prior.

[14]Obviously, the numerical standard errors are only (possibly rough) estimates of the actual stan-

Table 3: Sampling results for different candidate densities: estimated posterior moments for the 2-regime mixture AR(2) model (23) (with $\beta_{11} < \beta_{21}$) for (annualized) quarterly real US GNP growth in 1950-2007. The IS and MH methods based on the Type 3 network, the mixture of Student's t densities, yield much more precise results in approximately the same computing time.

	importance function / candidate density					
	Type 3 NN (Adaptive Mixture of t [AdMit])		Student's t with adapted mode and scale		Student's t around posterior mode	
	IS	MH	IS	MH	IS	MH
$E(\beta_{11})$	0.1292	0.1403	0.2022	-0.0091	0.2519	0.0527
(num. std. error)	(0.0060)	(0.0080)	(0.1191)	(0.1398)	(0.1049)	(0.0883)
[RNE]	[0.0058]	[0.0031]	$[5.9 \cdot 10^{-6}]$	$[7.3 \cdot 10^{-6}]$	$[6.2 \cdot 10^{-6}]$	$[1.6 \cdot 10^{-5}]$
$E(\beta_{12})$	0.4061	0.4044	0.4135	0.4518	0.3387	0.2225
(num. std. error)	(0.0030)	(0.0036)	(0.0157)	(0.0124)	(0.0367)	(0.1564)
[RNE]	[0.0092]	[0.0064]	[0.0001]	[0.0004]	$[2.6 \cdot 10^{-5}]$	$[5.9 \cdot 10^{-6}]$
$E(\beta_{13})$	0.2663	0.2606	0.2006	0.3303	0.2080	0.5178
(num. std. error)	(0.0037)	(0.0047)	(0.0705)	(0.0762)	(0.0807)	(0.1190)
[RNE]	[0.0050]	[0.0029]	$[5.8 \cdot 10^{-6}]$	$[7.3 \cdot 10^{-6}]$	$[4.8 \cdot 10^{-6}]$	$[4.9 \cdot 10^{-6}]$
$E(\beta_{21})$	1.7832	1.8398	1.3947	1.0742	0.9180	0.7467
(num. std. error)	(0.0154)	(0.0303)	(0.1186)	(0.0989)	(0.0517)	(0.0568)
[RNE]	[0.0044]	[0.0012]	$[6.1 \cdot 10^{-6}]$	$[8.0 \cdot 10^{-6}]$	$[9.3 \cdot 10^{-6}]$	$[1.2 \cdot 10^{-5}]$
$E(\beta_{22})$	-0.2569	-0.2546	-0.5583	-0.1474	-0.4611	0.0867
(num. std. error)	(0.0062)	(0.0108)	(0.2090)	(0.2045)	(0.1786)	(0.1800)
[RNE]	[0.0057]	[0.0020]	$[3.1 \cdot 10^{-6}]$	$[3.1 \cdot 10^{-6}]$	$[3.0 \cdot 10^{-6}]$	$[3.0 \cdot 10^{-6}]$
$E(\beta_{23})$	-0.0852	-0.0999	0.0986	0.0294	0.6317	0.1591
(num. std. error)	(0.0089)	(0.0098)	(0.0644)	(0.0823)	(0.1831)	(0.2118)
[RNE]	[0.0022]	[0.0018]	$[3.1 \cdot 10^{-6}]$	$[2.8 \cdot 10^{-6}]$	$[2.8 \cdot 10^{-6}]$	$[1.5 \cdot 10^{-6}]$
$E(\sigma)$	0.8367	0.8367	0.8239	0.8304	0.8253	0.8246
(num. std. error)	(0.0006)	(0.0008)	(0.0080)	(0.0049)	(0.0018)	(0.0011)
[RNE]	[0.0071]	[0.0034]	$[9.9 \cdot 10^{-6}]$	$[4.3 \cdot 10^{-5}]$	[0.0001]	[0.0006]
$E(p)$	0.6797	0.6887	0.7129	0.5380	0.7070	0.3753
(num. std. error)	(0.0039)	(0.0070)	(0.0941)	(0.1018)	(0.1032)	(0.0715)
[RNE]	[0.0068]	[0.0021]	$[4.3 \cdot 10^{-6}]$	$[4.7 \cdot 10^{-6}]$	$[3.1 \cdot 10^{-6}]$	$[7.5 \cdot 10^{-6}]$
$\sigma(\beta_{11})$	0.4605	0.4480	0.4091	0.5336	0.3698	0.5057
$\sigma(\beta_{12})$	0.2920	0.2891	0.2473	0.3454	0.2651	0.5348
$\sigma(\beta_{13})$	0.2577	0.2523	0.2394	0.2908	0.2509	0.3727
$\sigma(\beta_{21})$	1.0255	1.0491	0.4130	0.3957	0.2226	0.2830
$\sigma(\beta_{22})$	0.4723	0.4784	0.5209	0.5074	0.4357	0.4426
$\sigma(\beta_{23})$	0.4155	0.4103	0.1602	0.1957	0.4297	0.3657
$\sigma(\sigma)$	0.0474	0.0468	0.0358	0.0452	0.0254	0.0366
$\sigma(p)$	0.3213	0.3191	0.2748	0.3118	0.2572	0.2773
total time	185.2 s		196.5 s		188.5 s	
time construction NN	87.2 s					
time adapting mode, scale			8.0 s			
time sampling	98.0 s		188.5 s		188.5 s	
draws	1 mln		2 mln		2 mln	
time/draw	0.098 ms		0.094 ms		0.094 ms	

Again, about half of the computing time required for the results of IS or the MH algorithm using the AdMit candidate, is needed for 'learning' the candidate distribution, the mixture of 8 Student's t distributions. However, the RNE's are much higher for the AdMit candidate density, so that 1 million of AdMit draws are much more valuable than 2 millions of draws from a Student's t candidate distribution. Note that most RNE's are extremely low for the Student's t candidate distributions: for the parameter β_{11} it is approximately 6×10^{-6} under the IS approach, which means that the samples of 2 million draws are equivalent with a sample of merely 12 independent direct draws from the posterior! For the mixture of Student's t distributions, an RNE of 0.0058 may seem really low, as this means that the million draws are equivalent with merely 5800 independent direct draws. However, this mainly reflects that for highly non-elliptical posteriors in higher dimensions it may be almost impossible to (quickly) find a candidate distribution with a high RNE.

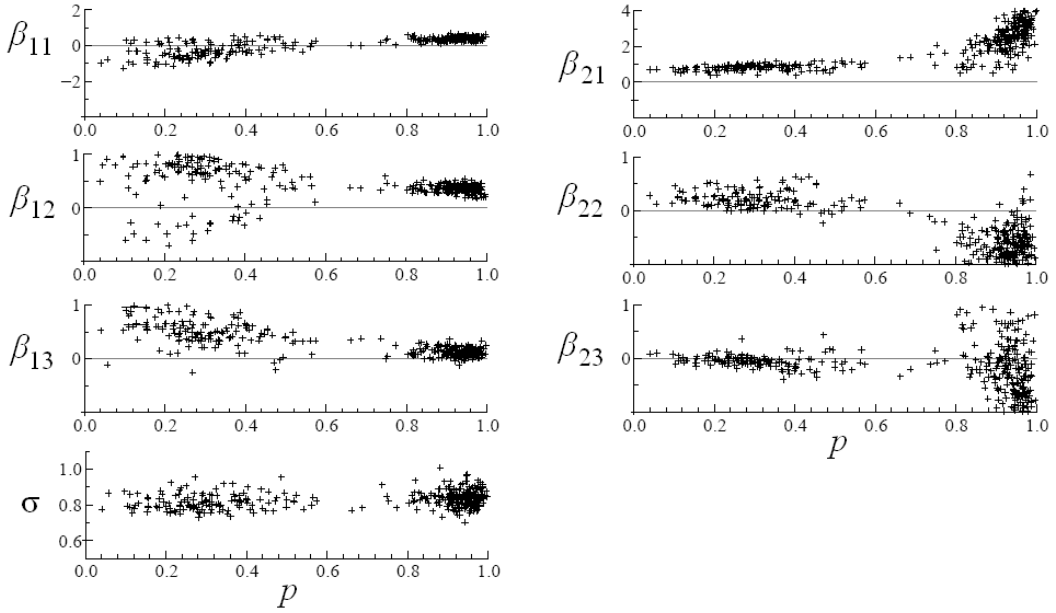
For both MH algorithms using Student's t distributions, there was a sequence of over 300000 consecutive rejections! This reflects that there are parts of the parameter space, which contain substantial posterior probability mass, that are almost completely 'missed' by these Student's t candidate distributions. This is illustrated by Figure 13. Another consequence is that some estimated posterior standard deviations are far too small for the Student's t candidate distributions. Figure 13 also shows that the posterior is bimodal. Further, Figure 13 reflects that if $p \to 0$ ($p \to 1$), then β_{11}, β_{12} and β_{13} (β_{21}, β_{22} and β_{23}) become unidentified, so that a wide range of values is possible for these parameters.

Finally, note that for the middle columns of Table 3 the mode and scale of the candidate have already been roughly adapted to the posterior, and that this is a fat-tailed Student's t distribution with 1 degree of freedom. Still substantial parts of the parameter space are almost completely missed, when using this unimodal candidate. This stresses the need for multi-modal candidate densities in such situations.

Of course, it is also possible to apply the method of Gibbs sampling with data augmentation to this 2-regime mixture model. However, our main aim is to compare the IS and MH algorithms that make use of different candidate distributions for an 8-dimensional, highly non-elliptical posterior distribution. Further, the data augmen-

dard deviations of the IS and MH estimators. This may explain the relatively large differences between the numerical standard errors under the IS and MH methods, and the smaller numerical standard errors (for some of the parameters) under the Student's t candidate distribution around the posterior mode as compared with the Student's t distribution with mode and scale adapted to the posterior.

Metropolis-Hastings draws (candidate = mixture of 8 Student's t distributions):



Metropolis-Hastings draws (candidate = Student's t distribution):

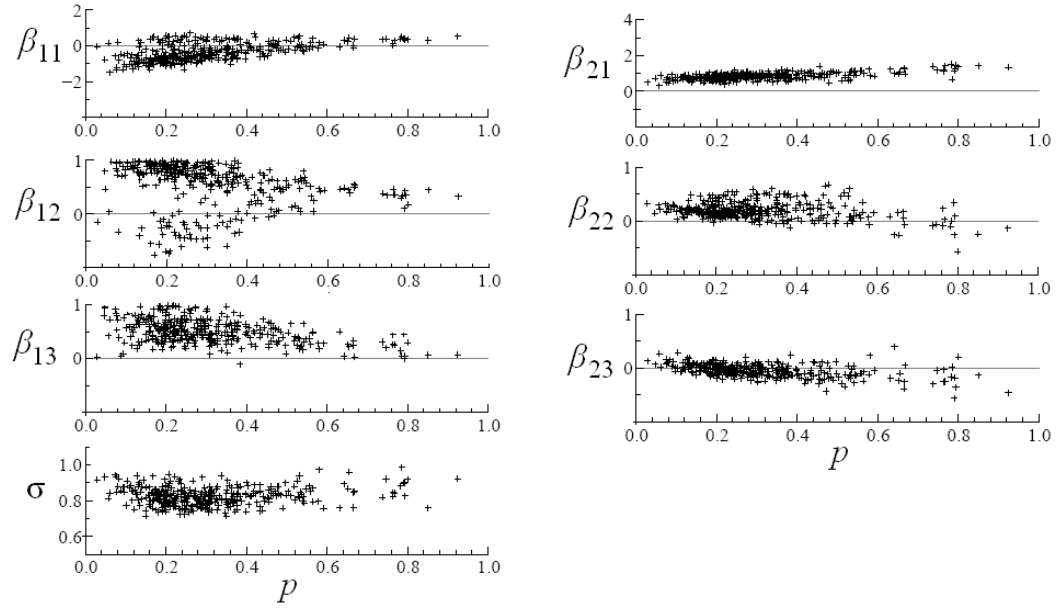


Figure 13: Posterior for the 2-regime mixture AR(2) model (23) (with $\beta_{11} < \beta_{21}$) for (annualized) quarterly real US GNP growth in 1950-2007: scatter plots of $\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23}, \sigma$ (on vertical axis) versus p (on horizontal axis) for Metropolis-Hastings draws with Adaptive Mixture of t [AdMit] candidate (top) and for Metropolis-Hastings draws with Student's t candidate (bottom). Notice that the regions with p close to 1 are (almost) completely 'missed' by the Student's t candidate distribution.

tation approach requires more 'inputs' than the IS and MH methods. For the data augmentation method, the conditional posterior distribution of each parameter has to be derived, whereas the IS and MH methods only require a kernel of the posterior density. In the case of multi-modality, the data augmentation approach may also fail, in the sense that the Gibbs sequence remains near one of the posterior modes. Obviously, one can then draw from the other regions of the parameter space by choosing a different initial value, but it is not a trivial issue how to weight the results from the different runs, i.e. it is not trivial to determine which part of the posterior probability mass is contained in each region of the parameter space.

Another approach is the permutation-augmented sampling method of Geweke (2007), which is close to the random permutation sampler of Frühwirth-Schnatter (2001). These approaches solve the problem of multimodality of the posterior in the unrestricted mixture model due to the symmetry of the mixture components in the (unrestricted) model. The idea behind the permutation-augmented approach is that one first generates draws from the unrestricted posterior and secondly permutes these in order to satisfy the identification constraint, where the second step is only performed if one desires insight into the *restricted* posterior distribution. However, the posterior distribution may also be highly non-elliptical 'per mode', which may cause slow convergence or unreliable results in case of high-dimensional posteriors. In such cases, a combination of the permutation-augmented idea and the mixture of Student's t distributions may be useful.

Finally, we briefly consider another irregularity of the likelihood of the mixture model that occurs if we allow the parameter σ to be different across regimes. Consider the simple mixture model:

$$y_t \sim \begin{cases} N(\mu_1, \sigma_1^2) & \text{with probability } p \\ N(\mu_2, \sigma_2^2) & \text{with probability } 1-p \end{cases} \qquad t = 1, \ldots, T, \tag{24}$$

where the y_t are independent. The likelihood function is:

$$\begin{aligned} L(\mu_1, \sigma_1, \mu_2, \sigma_2, p) \equiv p(y|\theta) \;=\; \prod_{t=1}^{T} \Bigg[& p\,(2\pi)^{-1/2}\,\sigma_1^{-1}\,\exp\left(-\frac{(y_t-\mu_1)^2}{2\sigma_1^2}\right) + \\ & (1-p)\,(2\pi)^{-1/2}\,\sigma_2^{-1}\,\exp\left(-\frac{(y_t-\mu_2)^2}{2\sigma_2^2}\right)\Bigg] \end{aligned} \tag{25}$$

with $y = (y_1, \ldots, y_T)'$, $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)'$. The likelihood function $L(\mu_1, \sigma_1, \mu_2, \sigma_2, p)$ in (25) has unbounded modes for $\mu_i = y_t, \sigma_i \to 0$ $(i = 1, 2;\ t = 1, \ldots, T)$, as for $\mu_i = y_t$ the factor $\exp\left(-\frac{(y_t-\mu_i)^2}{2\sigma_i^2}\right) = 1$, so that only the factor $\sigma_i^{-1} \to \infty$ remains.

Figure 14 shows the likelihood $L(\mu_1, \sigma_1, \mu_2, \sigma_2, p)$ for data on the (annualized) quarterly real US GNP growth rate y_t from the fourth quarter of 2005 to the third quarter of 2007 (the last 8 observations in Figure 12), conditional on the values $\mu_1 = 2.515$, $\sigma_1 = 1.478$, the mean and standard deviation of the 8 observations, and $p = 0.99$. Note the 8 'spikes' corresponding to the 8 observations y_t: we have $L(\mu_1 = 2.515, \sigma_1 = 1.478, \mu_2, \sigma_2, p = 0.99) \to \infty$ for $\mu_2 \to y_t$, $\sigma_2 \to 0$ $(t = 1, \ldots, 8)$.

This phenomenon means that the inverted gamma $IG(\nu_0, D_0)$ prior density for σ_i^2,

$$p(\sigma_i^2) = \frac{D_0^{\nu_0}}{\Gamma(\nu_0)} \left(\sigma_i^2\right)^{-\nu_0 - 1} \exp(-D_0/\sigma_i^2) \qquad D_0 > 0, \nu_0 > 0, \qquad i = 1, 2, \qquad (26)$$

used by e.g. Frühwirth-Schnatter (2001), can also be interpreted as a *regularization* prior in the sense that the exponent $\exp(-D_0/\sigma_i^2)$ eliminates the likelihood function's 'spikes'.

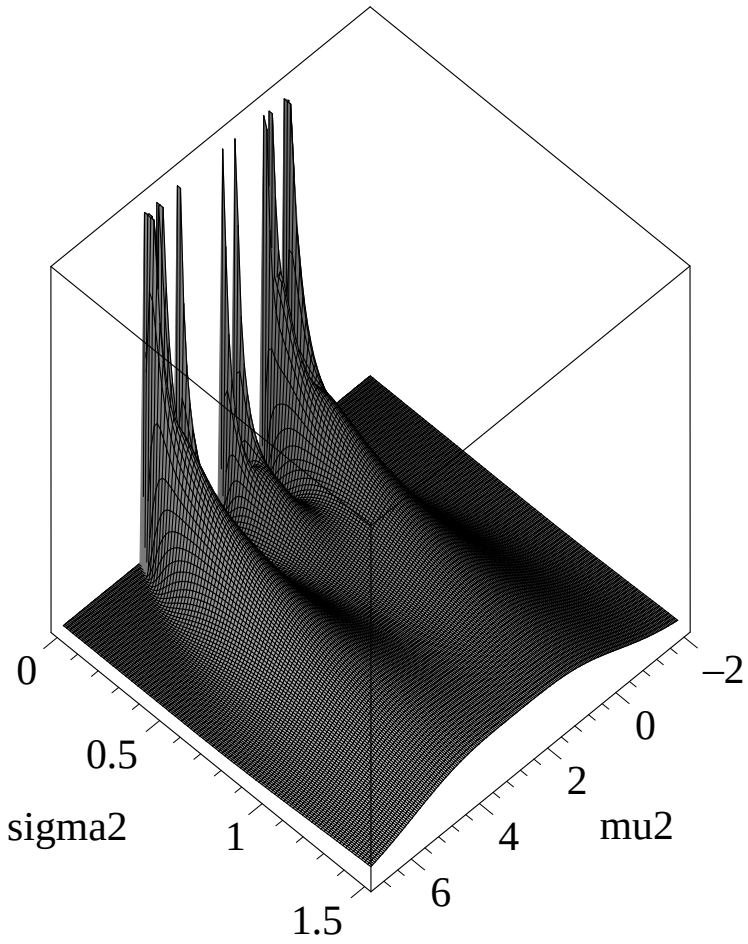


Figure 14: Likelihood function $L(\mu_1, \sigma_1, \mu_2, \sigma_2, p)$ in simple two-regime mixture model with different standard deviation parameters σ_1, σ_2 in the two regimes, for data on the (annualized) quarterly real US GNP growth rate y_t from 2005Q4 to 2007Q3 (the last 8 observations in Figure 12); conditional on the values $\mu_1 = 2.515$, $\sigma_1 = 1.478$, the mean and standard deviation of the 8 observations, $p = 0.99$. The likelihood $L(\mu_1 = 2.515, \sigma_1 = 1.478, \mu_2, \sigma_2, p = 0.99) \to \infty$ for $\sigma_2 \to 0$, $\mu_2 \to y_t$ $(t = 1, \ldots, 8)$.

7 Concluding remarks

In this paper, we considered the possibility of highly non-elliptical posterior distributions that may occur in several econometric models, in particular, when one allows the likelihood to dominate and the information in the data is weak. We investigated three cases: instrumental variable models with weak instruments, vector autoregressive models with co-integration restrictions, and mixture processes where one component is nearly non-identified.

We started with an analysis of the issue of highly non-elliptical posteriors in the context of a simple IV model for the effect of education on income using data from the well-known Angrist and Krueger (1991) study. We discussed how a so-called Information Matrix or Jeffreys prior may be used as a 'regularization prior' that in combination with the likelihood function yields posteriors with desirable properties. Further, we illustrated that the IV model and the Vector Error Correction Model have a similar mathematical structure which leads to similar posterior shapes.

As a main contribution of the paper, we find that in situations of highly non-elliptical posteriors that may occur frequently in economic processes, it is worthwhile to invest in the search for accurate candidate or importance functions. Simple simulation methods like the Metropolis-Hastings algorithm or Importance sampling with one normal or Student's t candidate density may either fail to converge or be extremely slow, which inhibits their use in practical applications. In all examples considered in this paper, the mixture of Student's t densities – that can be considered a particular type of neural network function – is clearly a much better candidate. This mixture candidate yields far more precise estimates of posterior means after the same amount of computing time, whereas the Student's t candidate almost completely misses substantial parts of the parameter space.

Of course, it is also possible to apply the method of Gibbs sampling with data augmentation to the 2-regime mixture model. However, our main aim is to compare the IS and MH algorithms that make use of different candidate distributions for an 8-dimensional, highly non-elliptical posterior distribution. Further, in the case of multi-modality, the data augmentation approach may also fail, in the sense that the the Gibbs sequence remains near one of the posterior modes.

Another approach is the permutation-augmented sampling method of Geweke (2007), which is close to the random permutation sampler of Frühwirth-Schnatter (2001). However, the posterior distribution may also be highly non-elliptical 'per mode', which may cause slow convergence in case of high-dimensional posteriors. In such cases, a

combination of the permutation-augmented idea and the mixture of Student's t distributions may be useful. The mixture of t candidate can also be applied to particular (non-linear) multivariate GARCH models, where application of the data augmentation method is more difficult. Another possible extension is the combination of copulas and mixtures of Student's t distributions, where the use of copulas helps the marginal candidate distributions match with the marginal posteriors. We intend to report on these extensions in the near future.

References

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59(3), 817–858.

Andrews, D.W.K., Monahan, J.C., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* 60(4), 953–966.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 91, 444–472.

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.

Berger, J.O., 1985. *Statistical Decision Theory and Bayesian Analysis*, Second Edition. Springer-Verlag, 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.

Brooks, S.P., Roberts, G.O., 1998. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing* 8, 319–335.

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.

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.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and

dynamic switching and mixture models, *Journal of the American Statistical Association* 96, 194-209.

Gallant, A.R., White, H., 1988. There exists a neural network that does not make avoidable mistakes. 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.

Geweke, J., 2007. Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics & Data Analysis* 51, 3529–3550.

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

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, 2006. Essays on Neural Network Sampling Methods and Instrumental Variables. Ph.D. thesis, Tinbergen Institute, Erasmus University Rotterdam.

Hoogerheide, L.F, Van Dijk, H.K., 2006. A reconsideration of the Angrist-Krueger analysis on returns to education. Econometric Institute report EI 2006-15, Erasmus University Rotterdam.

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

Hoogerheide, L.F, Kleibergen, F.R., Van Dijk, H.K., 2007. Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. *Journal of Econometrics*, 138(1), 63–103.

Hoogerheide, L.F., Van Dijk, H.K., Van Oest, R.D., 2008. Simulation Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances. Chapter in *Handbook of Computational Econometrics*, Elsevier, forthcoming.

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

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

Imbens, G.W., Angrist, J.D., 1997a. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 25, 305–327.

Imbens, G.W., Angrist, J.D., 1997b. Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies* 64, 555–574.

Kleibergen, F.R., Van Dijk, H.K., 1994. 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.

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.

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.

Siegel, P., Hodge, R., 1968. A Causal Approach to the Study of Measurement Error. In: Blalock, H., Blalock, A. (eds.), *Methodology in Social Research*. McGraw-Hill, New York.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82, 528–550.

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

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 im-

portance sampling in Monte Carlo integration. In: Bernardo, J.M., Degroot, M., Lindley, D. and Smith, A.F.M. (Eds.), *Bayesian Statistics 2*, Amsterdam, North-Holland.

White, H., 1989. Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association* 84, 1003–1013.

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.