ESSAYS ON NEURAL NETWORK SAMPLING METHODS AND INSTRUMENTAL VARIABLES

ISBN 00 0000 000 0

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is number **379** of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Essays on Neural Network Sampling Methods and Instrumental Variables

Essays over neurale netwerken in simulatie methoden
en instrumentele variabelen

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. S.W.J. Lamberts
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
donderdag 29 juni 2006 om 13.30 uur

door

Lennart Frank Hoogerheide
geboren te Rotterdam

Promotiecommissie

Promotor:	Prof.dr. H.K. van Dijk
Overige leden:	Prof.dr. L. Bauwens
	Prof.dr. P.H.B.F. Franses
	Dr. R. Paap

Acknowledgements

Many persons have contributed to the realization of this thesis. First of all, I would like to thank my promotor Herman van Dijk for his support and enthusiasm in the past years; his new ideas and optimism have always been a major encouragement. Additionally, credits should go to my other coauthors Johan Kaashoek, Frank Kleibergen and Rutger van Oest. I greatly benefitted from the cooperation with them. Next, I wish to express my gratitude to Philip Hans Franses and Richard Paap for useful comments on an earlier draft of this thesis.

In particular, I would like to mention my appreciation for Rutger van Oest, who shared a room at the Tinbergen Institute with me for years and still did not lose his sanity, for many entertaining moments. I also want to thank my (former) roommates at the Econometric Institute Govert Bijwaard, Max Ekhorst and Björn Vroomen, and the other (former) colleagues at the Econometric and Tinbergen Institutes for providing a good atmosphere, amongst whom I would like to explicitly mention Paul de Boer, Jan Brinkhuis, Bram van Dijk, Dick van Dijk, Bas Donkers, Felix Eschenbach, Dennis Fok, Philip Hans Franses, Patrick Groenen, Christian Hafner, Christiaan Heij, Johan Kaashoek, Richard Klein, Alex Koning, Martin Martens, Martyn Mulder, Richard Paap, Michiel de Pooter, Francesco Ravazzolo, Joost van Rosmalen, René Segers, Klaas Staal, Roel Stroeker, Albert Wagelmans and Phongtorn Wrasai. Further, I would like to thank the staff at the Econometric and Tinbergen Institutes; especially, I want to mention my appreciation for Aletta Henderiks, Carine Horbach, Carien de Ruiter and Dave van Rutten who also substantially contributed to a nice working atmosphere.

An important part of working at the Econometric Institute consisted of giving lectures and co-supervising bachelor and master theses. I would like to thank those students who made this work enjoyable. I especially enjoyed the co-supervision of the master thesis of Remco Rothkrantz.

While working on my dissertation, I had the opportunity to present my research at several conferences in France, Italy, Japan, the Netherlands, Switzerland and the United

States. Financial support from the Econometric Institute, the Tinbergen Institute, SAMSI and the Tokyo Metropolitan University is gratefully acknowledged. Further, I would like to thank Luc Bauwens for inviting me to present a seminar at CORE.

Finally, I thank my family for their moral support, for showing interest and always being available when needed. In this respect, I particularly would like to thank my parents, my sister, Martijn, Vrijdag and Merel.

Lennart Hoogerheide
Rotterdam, May 2006.

Contents

Chapter 1

Introduction

1.1 Motivation and structure of the thesis

In this thesis new results are given for instrumental variables (IV) regression models, and a class of sampling methods is introduced and explored. These sampling methods, which make use of neural network approximations to posterior densities, can quickly simulate draws from posterior distributions in many models. In this section the relevance of the research is stressed, and the structure of the thesis is explained.

1.1.1 Instrumental variables

Measuring the effect of education on income, the (monetary) return on education, 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, smarter 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.[1] Another problem is the measurement error in reported education. First, usually only the completed number of years of education is reported. Second, people may misreport their education spell.[2] 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. 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.

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. This data set contains the number of completed years of education and the logarithm of weekly earnings in 1979. Figure 1.1 illustrates the difference between simply regressing income on education and using quarter of birth to form an instrumental variable. The left panel shows

[1] The intellectual capabilities of the persons in the sample may not be the only reason for an overestimated effect of education on income. The (often unobserved) intellectual capabilities, income and education level of their parents may also cause an upward bias, as these characteristics of their parents may also influence their education level and have a direct effect on their income; for example, it may be the case that children of more intelligent and higher educated parents on average learn more at home.

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

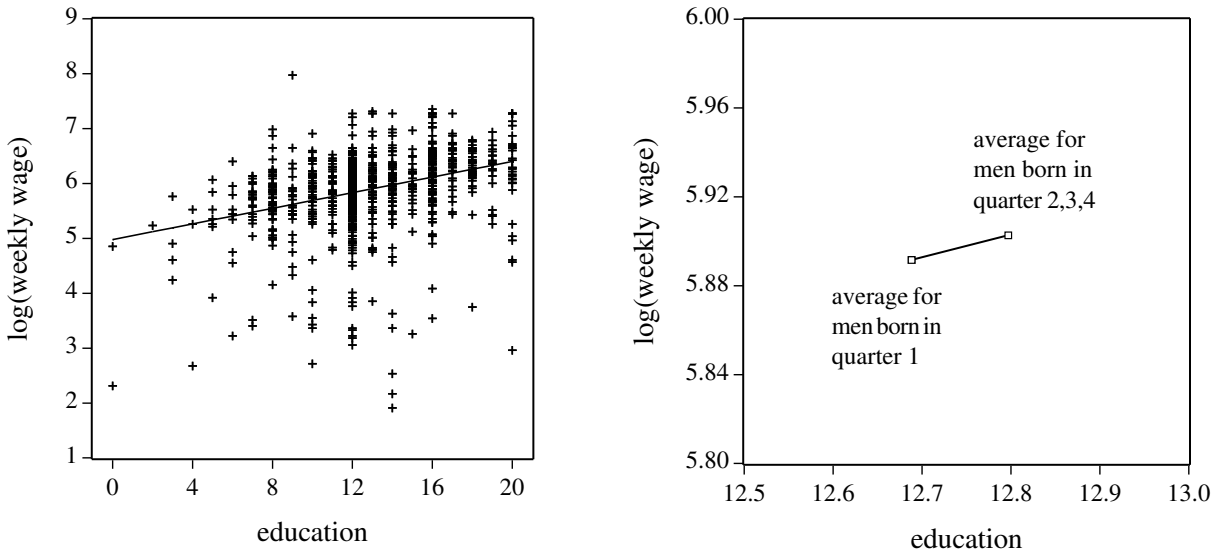

Figure 1.1: Measuring the effect of education on income: simple regression of (logarithm of) income on education (left), or using quarter of birth as an instrumental variable (right)

how the effect of education on income is estimated by simple regression. The estimate is the steepness of the regression line, the line which minimizes the sum of squared (vertical) deviations of points from this line. For all data of the US this estimate is 0.0709: each added year of education results on average in a 7.09% increase in income. However, this method may overestimate or underestimate the effect of education on income because of latent intellectual capabilities or measurement errors, respectively. The right panel illustrates how the effect of schooling on earnings can be estimated using quarter of birth as an instrument. The average education spell for men born in the first quarter is 12.6881 years, while for men born in other quarters the average education spell is 12.7969. So, the schooling laws imply that men born in the first quarter on average have 0.1088 years less education than men born in the other quarters. Further, for men born in the first quarter the average logarithm of income is 0.0111 ($= 5.9027 - 5.8916$) less than for those born in other quarters. In other words, men born in the first quarter have on average an income that is (approximately) 1.11% lower than men born in the other quarters. The key assumption is now that quarter of birth only influences income because of its effect on education, so that we may interpret the 1.11% difference in average income as a result of the difference in average education spell of 0.1088 years: each added year of education results on average in a 10.20% ($= 0.0111/0.1088$) increase in income. So, at first sight

it seems that if any bias exists in the simple regression, then this is a downward bias: measurement errors in reported years of education may have caused an underestimation of the return on education. However, in the abovementioned approaches we have only obtained estimates of the effect of education spell on income, but we have no measure of the uncertainty on these estimates: we have no lower and upper bounds between which the effect of education on income lies (with a certain probability). In order to obtain such a probability interval we must specify a model; this will be done in the sequel of this section.

First note that if the average education spell is exactly the same for those born in the first quarter and the others, then the approach using quarter of birth as an instrument does not work. In that case one can not identify the difference in income per year of education, as this leads to a division by zero. This illustrates that in instrumental variables models the problem of *local non-identification* may occur: if the instrument (quarter of birth) has no effect on the explanatory variable (education), then one can not identify the effect of the explanatory variable on the variable that is to be explained (income) using this instrument. Furthermore, if the average education spell is almost equal for those born in the first quarter and the others, then there is obviously much uncertainty on the estimated return on education. For in this case some changes of education and/or income for a few persons would result in a quite different estimate of the return on education. This kind of situation in which instruments only explain a small fraction of the variation in (some of) the explanatory variables, is usually referred to as the case of *weak instruments*. In fact, the difference in average education spell of 0.1088 years is small as compared to the variation in education spells across individuals (with education spells varying between 0 and 20 years, having a standard deviation of 3.28 years), so that the uncertainty on the estimate of the return on education is obviously much larger in the approach using quarter of birth as an instrument than in the simple regression. In other words, much information is lost by merely using the averages of education and income for the two quarter-of-birth groups; in the extreme case where the average education spell would be exactly the same for both groups, no information on return on education would be left. So, although the systematic error (of over- or underestimation) due to latent capabilities or measurement errors is avoided by using instruments, the weakness of the instruments may cause probability intervals for the return on education that are so wide that they are of little practical use.

Before specifying a model, it should be noted that the main approach of making inference in this thesis is the Bayesian approach. Only in chapter 5 the results from Bayesian methods will be compared with results from classical approaches. In the classical approach probabilities are objective: probability is defined as the fraction of occurrences, or frequency, when a process is repeated infinitely often. Hence, the classical approach is also known as the frequentist approach. In the classical framework, testing is based on comparing the value of a test statistic (that is computed using the observed data) with the corresponding distribution of the test statistic resulting from an infinitely large hypothetical data set from a certain 'true' data generating process. If the realization of the test statistic is unlikely for the assumed 'true' data generating process, the underlying hypothesis is rejected. In the classical approach, parameters are considered as unknown constants that have to be estimated, typically using the method of maximum likelihood.

In the Bayesian approach the parameters of a model are considered as random variables. The prior density of these parameters, reflecting one's prior beliefs before observing data, is updated by the likelihood function, reflecting the information in the data, resulting in the posterior density. In the Bayesian framework probabilities are subjective: probability distributions reflect beliefs which may differ between persons. In the Bayesian framework no assumption is made that one can hypothetically repeat the process of interest infinitely many times.

A possible advantage of Bayesian analysis over classical inference is that it may be easier to assess the uncertainty of results in the Bayesian framework, especially in small samples. For example, the distribution of the maximum likelihood estimator depends on the 'true' values of the parameters which are unknown. In the Bayesian framework one specifies a prior density for the parameters, after which a kernel proportional to the posterior density is given by the prior multiplied with the likelihood function. If this kernel corresponds to a proper density function in the sense that it integrates to a finite number, one can use integration methods in order to evaluate characteristics such as standard deviations or intervals containing 95% of the probability mass for the posterior distribution of the parameter(s) of interest.

We now consider a simple, illustrative model for the returns to schooling. First define $D_{quarter,i}$ as the following 0/1 variable: $D_{quarter,i} = 1$ if person i is born in quarter 2,3 or

4, and $D_{quarter,i} = 0$ if person i is born in quarter 1. The model is as follows:

$$\log wage_i = \beta\, education_i + \varepsilon_i \tag{1.1}$$

$$education_i = \pi\, D_{quarter,i} + v_i \tag{1.2}$$

for $i = 1, 2, \ldots, T$, where $\log wage$, *education* and $D_{quarter}$ are taken in deviation from their means, so that no constant terms occur in (1.1) and (1.2). 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 100β %. The 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. This is the case of *exact identification* in which there are as many instruments (only $D_{quarter}$) as explanatory endogenous variables (only *education*). The error terms ε_i and v_i are assumed to be independent across observations and normally distributed:

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim N(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

We specify the following non-informative prior density kernel of Drèze (1976):

$$p(\beta, \pi, \Sigma) \propto |\Sigma|^{-h/2} \text{ with } h > 0. \tag{1.3}$$

Drèze (1976, 1977) gives the joint posterior kernel of (π, β) and the marginal posterior kernel of β that follow from the prior specification in (1.3). Figure 1.2 shows the shapes of the joint posterior kernel of (π, β) and the marginal posterior kernel of β (for the choice of $h = 3$) on bounded domains for several data sets: data of all states of the US, data of the state of New York, and data of the states of Kentucky, Tennessee and Arkansas.

First it should be noted that in this case of exact identification, both the joint posterior of (π, β) and the marginal posterior of β under the flat prior (1.3) are improper on $\mathbb{R}^2$ and $\mathbb{R}$, respectively: the integrals of the joint and marginal posterior density kernel are infinite. This results since in the joint posterior kernel of (π, β) a nonintegrable asymptote is present around $\pi = 0$. Although improper on $\mathbb{R}^2$, the joint posterior of (π, β) can be made proper by restricting β and/or π to a certain area. In that case it depends on the data y, x and z, whether the behavior for $\pi = 0$ still dominates the analysis. For example, for data of all states of the US and for data of Kentucky, Tennessee and Arkansas the joint posterior of (π, β) has a clear peak away from $\pi = 0$. This indicates that a sufficiently large difference in average education spell exists between men born in the first quarter and the others, so that valuable results on the returns to schooling can be obtained. In these cases the marginal posterior of β seems to have a bell-shape (with a peak around

$\beta = 0.10$). On the other hand, for data of the state of New York the joint posterior kernel (π, β) displays a ridge around $\pi = 0$. In New York there is no or little difference in average education spell between the two quarter-of-birth groups, so that the instrumental variable (IV) approach gives no or little information on returns to schooling. The parameter π can take values close to 0, and for these values of π the parameter β can take a wide range of values; this reflects the local non-identification of β for $\pi = 0$. This leads to a marginal posterior of β with fat tails. Notice that the data set of New York even has somewhat more observations than the data set of Kentucky, Tennessee and Arkansas, so that it is not the size of the data set that causes the difference in the posterior of β. The only reason is the huge difference in strength of the quarter-of-birth instrument between the states. In chapter 5 of this thesis a more advanced model for the returns to schooling is considered. It allows for differences in average education spell between all quarters (instead of merely permitting a difference between the first quarter and the rest). Furthermore, these differences in education between quarters are allowed to differ between years and states of birth. Also a direct effect of state and year of birth on income is allowed. This model is also considered by Angrist and Krueger (1991), who conclude that there is little bias in the ordinary least-squares (OLS) estimate and that, if anything, the conventional OLS estimate is biased slightly downward. Note that the latter corresponds with the results from the simple regression and the instrumental variables approach in the aforementioned simple example. Angrist and Krueger (1991) only apply a classical method, two-stage least-squares (2SLS), and consider only results for the whole data set of all states of the US. In chapter 5 of this thesis the research by Angrist and Krueger (1991) is extended. Results are examined for both classical and Bayesian methods. Furthermore, it is considered how results vary between subsets of the data corresponding to regions of the US. We divide the US into four regions that are also used by the US Census Bureau, the source of the data. The states and numbers of observations in each region are given by Table 1.1. The strength of the quarter-of-birth instruments greatly differs between the Census regions; this can be seen from Figure 1.3, which shows for the four Census regions the marginal posterior of β, the effect of education on income, under a non-informative prior, the Jeffreys prior.[3] The marginal posterior distribution of β for the US under the Jeffreys prior is almost completely determined by the region South; the difference between the posterior of β for the US and for the South is small, whereas the 95% posterior density intervals are relatively large for the other regions. This indicates that the quarter-of-birth

[3]For the derivation of (an accurate approximation of) the marginal posterior of β under the Jeffreys prior, the reader is referred to Kleibergen and Zivot (2003).

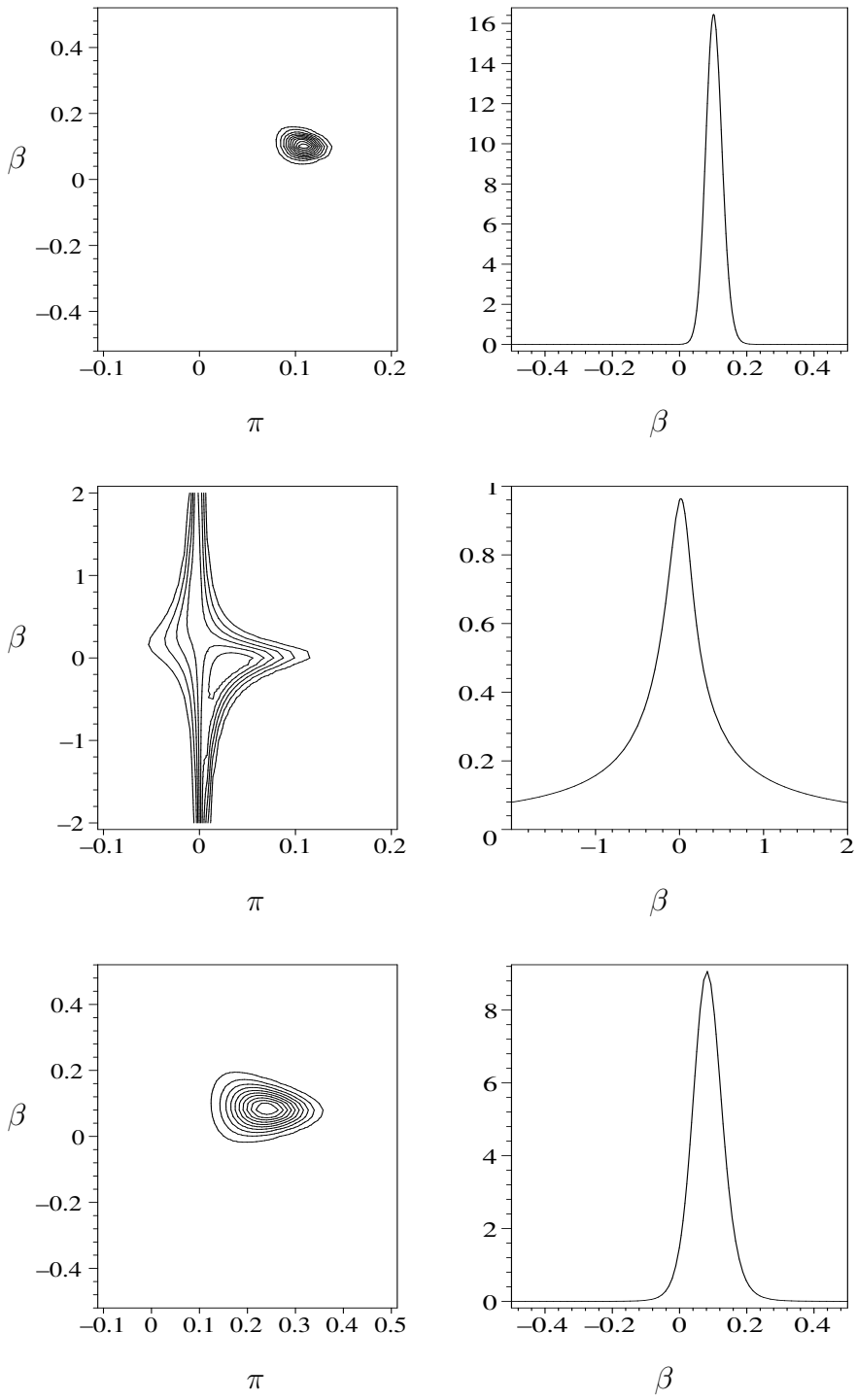

Figure 1.2: Contour plot of joint posterior density kernel of (π, β) (left) and marginal posterior density kernel of β (right) for data of US (top, $T = 329509$) the state of New York (middle, $T = 29015$), the states of Kentucky, Tennessee and Arkansas (bottom, $T = 23062$)

Table 1.1: US Census Bureau Regions

	Census region	number of observations	number of states	states (including D.C)
1.	Northeast	84484	9	Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont.
2.	Midwest	102267	12	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin.
3.	South	114391	17	Alabama, Arkansas, Delaware, D.C., Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia.
4.	West	28367	13	Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming.
	USA	329509	51	

instruments are much stronger in the region South than in the other regions.[4] This is illustrated by Figure 1.4, which reflects the results of the multiple F-test in the *first stage regression*, the regression of education on dummy variables that are based on the quarter of birth, for data of one state. The three states with p-value smaller than 0.001, Arkansas, Kentucky and Tennessee, are neighboring states in the region South. If the effect of the return on education is different for the other regions, which can not a priori be ruled out given the large economic differences between these regions, inference using data of the US is not representative for the average returns on education across the US. One should therefore be careful when drawing such conclusions.

In chapter 5 the well-known criticism of Bound, Jaeger and Baker (1995) is also considered. Bound, Jaeger and Baker (1995) have concluded that the interaction between compulsory school attendance laws and quarter of birth, which is the basis of the models of Angrist and Krueger (1991), does not give much usable information concerning the causal effect of education on wages for two main reasons. First, the weakness of the instruments may lead to large inconsistencies in the 2SLS estimator even if there is only a weak relationship between the instruments and the error in the structural equation; Bound, Jaeger and Baker (1995) mention evidence casting doubt on the assumption that no such correlation is present. Moreover, Bound, Jaeger and Baker (1995) even report that differences in family income at time of birth would seem to account for virtually all of the association between quarter of birth and wages: they argue that the difference in

[4]It is shown in Hoogerheide, Kleibergen and Van Dijk (2006) that Bayesian analysis using the Jeffreys prior, similar to the limited information maximum likelihood (LIML) estimator, focusses on the strongest available instruments.

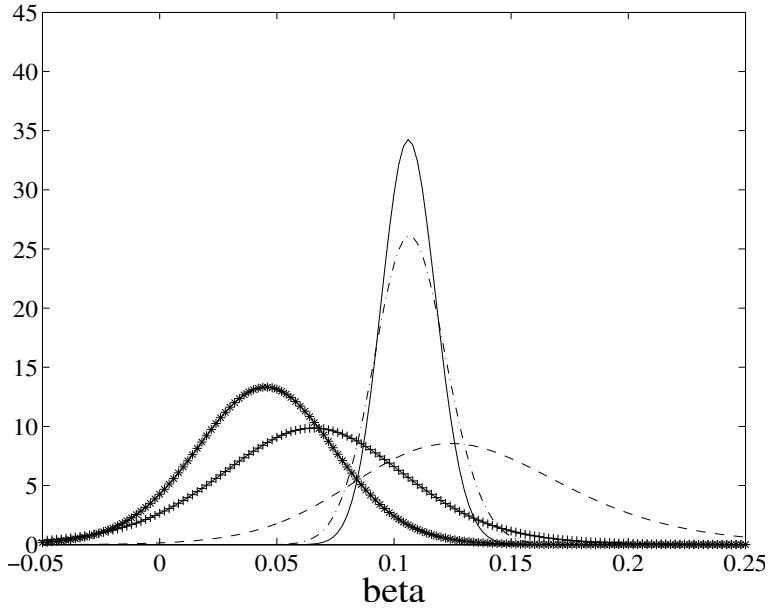

Figure 1.3: Marginal posterior of return on education β under Jeffreys prior for US (solid), Northeast (solid-plusses), Midwest (dashed), South (dash-dot), West (solid with stars).

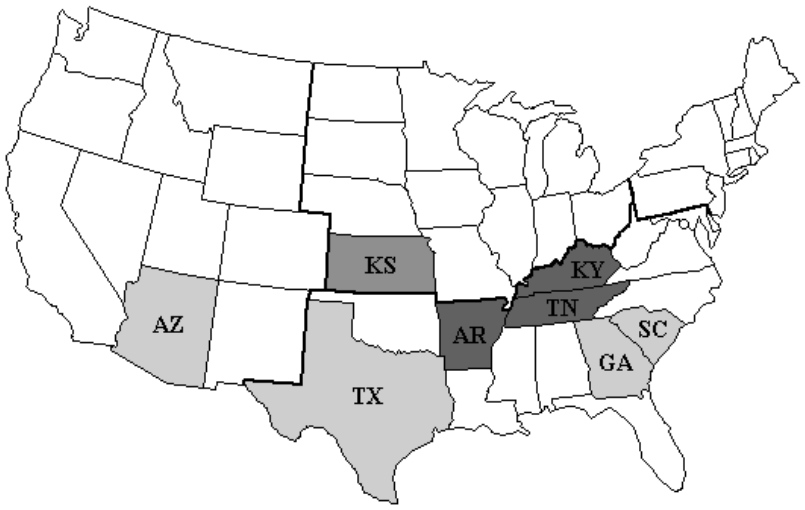

Figure 1.4: p-value of multiple F-test in first stage regression, regression of education on dummy variables that are based on quarter of birth, for data of individual states: p-value < 0.001: dark grey, p-value < 0.01: grey, p-value < 0.1: light grey.
(AR = Arkansas, AZ = Arizona, GA = Georgia, KS = Kansas, KY = Kentucky, SC = South Carolina, TN = Tennessee, TX = Texas)

income between those born in the first quarter and those born during the rest of the year can almost completely be explained by differences in family income at time of birth and an intergenerational correlation. Second, the 2SLS estimates reported by Angrist and Krueger (1991) may suffer from substantial finite sample biases because of the weakness of the instruments (despite the large sample size).

However, the second problem can be solved by using the Bayesian approach under the Jeffreys prior instead of the classical 2SLS method; or another alternative to the 2SLS estimator is the limited information maximum likelihood (LIML) estimator which is approximately median unbiased in this case. Furthermore, it is shown in chapter 5 that one may still obtain a rather tight posterior for β if one allows for a direct effect of birth during the first quarter on income. So, it seems that the conclusion of Bound, Jaeger and Baker (1995) concerning the use of instruments based on the quarter of birth is too strong, as a model of Angrist and Krueger (1991) (or a slightly modified version) can give usable information on the causal effect of education on income in (regions of) the US.

The difference between the posteriors for data of New York and data of Kentucky, Tennessee and Arkansas in Figure 1.2 illustrates that in instrumental variable regression models the shape of the posterior density may greatly differ between cases of relatively weak and strong instruments. In chapter 4 of this thesis, it is shown in a systematic manner how the shapes of posteriors depend on both the strength of the instrumental variables and the level of endogeneity under two diffuse prior specifications: the flat prior of Drèze (1976, 1977) and the Jeffreys prior. Bayesian analysis of the instrumental variable (IV) regression model under the flat prior has the following two peculiar properties. First, the joint posterior of (π, β) has an asymptote at $\pi = 0$, which is nonintegrable in the case of exact identification. Second, the tail behavior of the marginal posterior of β depends on the number of instruments in the sense that its tails become thinner when (possibly irrelevant) instruments are added to the model.[5] In the IV regression model, the Jeffreys prior can be considered as a 'regularization prior' in the sense that it 'remedies' these two peculiar properties of the posterior under the flat prior. In chapter 4 of this thesis, the effect of the Jeffreys prior is examined and illustrated in a systematic way; although the Jeffreys prior remedies certain peculiar properties, it may still lead to highly non-elliptical posteriors.

[5]This result appeared in an informal way in Maddala (1976), commenting on Drèze (1976).

1.1.2 Neural network sampling methods

Evaluating integrals is a crucial ingredient in Bayesian inference. The reason is that one starts with a kernel of the joint posterior density of all parameters occurring in the model, which is obtained by multiplying the prior density kernel with the likelihood function, whereas one is typically interested in the posterior means and standard deviations of (some of) the parameters, or in the posterior probability that a parameter lies in a certain interval. For these purposes one has to integrate the joint posterior density kernel with respect to all parameters.

The range of models and prior densities for which the integration of the joint posterior density kernel can be performed analytically is very restricted. In many cases numerical integration methods are required. Basically there are two numerical approaches: deterministic integration and Monte Carlo integration. Deterministic integration consists of evaluating the integrand at a set of many fixed points, and approximating the integral by a weighted average of the function evaluations. Monte Carlo integration is based on the idea that the posterior mean of a parameter can be approximated by the sample mean of a set of draws of the parameter from the posterior distribution. At a first glance, deterministic integration may always seem a better idea than Monte Carlo integration, as no extra uncertainty (caused by the required random variables) is added to the procedure. However, in deterministic integration the number of required function evaluations increases exponentially with the dimension of the integration problem k, the number of parameters over which one has to integrate. Therefore, deterministic integration approaches like quadrature methods become unworkable if k exceeds, say, three. So, in many cases one has to make use of Monte Carlo integration. However, for many models (and prior densities) it is impossible to directly draw from the posterior distribution. Then one has to use indirect sampling algorithms, for example importance sampling or the Metropolis-Hastings algorithm.

Importance sampling, due to Hammersley and Handscomb (1964), was introduced in econometrics and statistics by Kloek and Van Dijk (1978). Roughly speaking, importance sampling consists of drawing a set of points from a candidate density, also known as the importance function, and approximating the mean of the (posterior) distribution of interest by the weighted average of the sampled values, where the weights (adding to one) are proportional to the ratio of the 'target' (posterior) density and the candidate density. In this way the sample of points drawn from the candidate is 'corrected': points in regions (of the parameter space) where the candidate is low and the posterior is high get high

weights, whereas points in regions where the target (posterior) density is neglectable, get neglectable weights and are practically deleted from the sample.

The Metropolis-Hastings (MH) algorithm is a Markov chain Monte Carlo (MCMC) approach that has been introduced by Metropolis et al. (1953) and generalized by Hastings (1970). Markov chain Monte Carlo methods construct a Markov chain converging to a target distribution, in our case the posterior distribution of interest. After a burn-in period, which is required to make the influence of initial values negligible, draws from the Markov chain are considered as (correlated) draws from the target distribution itself. In the independence chain MH algorithm a candidate draw is sampled independently from the current state. The candidate draw is either accepted in which case the next state in the Markov chain is the candidate draw, or rejected in which case the next state in the Markov chain is the same as the current state.[6] In this way the sample of points drawn from the candidate density is 'corrected': points in regions (of the parameter space) where the candidate density is low and the (posterior) target density is high, are repeated several times in the Markov chain (as they are accepted themselves after which several other candidate draws are rejected). On the other hand, points in regions where the posterior is neglegible are always rejected: these points are removed from the sample.

In a 'standard' case of importance sampling or the independence chain MH algorithm, the candidate distribution is unimodal; a common choice is a normal or Student-t distribution. This 'standard' approach works well for (nearly) elliptical target distributions, unimodal distributions that are 'close' to the normal or Student-t distribution, for which in the two-dimensional case the lines in a contour plot, consisting of points for which the target density is equal, are (almost) ellipses. However, if the target (posterior) distribution is very different from the normal or Student-t distribution, the convergence behavior of these 'standard' Monte Carlo integration methods is rather uncertain. For example, consider the joint posterior of π and β for data of New York in Figure 1.2 (where β is restricted to lie in the interval [-2,2]; without restricting the domain of β (and/or π) the integral of the joint posterior kernel is infinite).[7] If we now use a tight normal distribution around the posterior mode (at $\pi \approx 0.03, \beta \approx -0.1$) as the candidate distribution in IS or

[6]The probability that the candidate draw is accepted depends on the ratio of the target density and the candidate density. If this ratio is larger at the candidate draw than at the current state, then the candidate draw is always accepted. Otherwise, the probability is given by the ratio of these ratios.

[7]It should be noted that in this case one can analytically integrate with respect to π. Furthermore, deterministic integration would be preferable to Monte Carlo integration in this 2-dimensional case. However, as indicated above, these methods are in general impossible (or too slow) in higher dimensions, so that a sampling method such as IS or MH is required.

the MH algorithm, then parts of the ridge around $\pi = 0$ with (in absolute sense) large values of β may be missed. In that case the uncertainty on β, the effect of education on income, is underestimated. Then IS or the MH method yields an estimate of the 95% posterior density interval of β that is far too tight, so that there is a large probability that this interval does *not* contain the true effect of education on income. In other words, the specification of a poor candidate density may lead to the incorrect conclusion that there is much information on the returns on education present in the data of New York, even though these data contain hardly any information on the causal effect of schooling on income. Of course, one could solve this problem by increasing the scale of the normal candidate distribution: if one chooses a large enough standard deviation for β, one also samples (in absolute sense) large β values. However, increasing the scale of the normal candidate distribution would imply that many points are drawn in areas with negligible posterior probability mass; for example, in our case such regions are the areas around the points $(\pi, \beta) = (0.05, 1)$ and $(\pi, \beta) = (0.05, -1)$. Further, (relatively too) few points are drawn close to the mode; these points therefore get huge IS weights. These disadvantages imply that either conclusions are unreliable or at least many draws are required, when using a normal or Student-t candidate distribution in the case of highly non-elliptically shaped posteriors. In the latter case enormous amounts of computing time may be required, especially in certain high-dimensional highly non-elliptical distributions, even on a modern computer. This may be a problem in many situations, especially in certain financial applications where computations should be 'real time', as quick decisions are often required. For example, when deciding whether or not to adjust a portfolio on the basis of a Bayesian analysis of a complex non-linear model, a difference in computing time between 15 minutes and 2 hours may be very relevant. So, for highly non-elliptical target distributions one mostly needs to look for a more appropriate candidate distribution than a normal or Student-t distribution in IS or the MH algorithm (or for a different Monte Carlo integration method). Geweke (1989) stresses that in general for quick convergence of the sampling results the candidate density should be 'close' to the target density, and that it is important that the tails of the candidate distribution should not be thinner than those of the target.

One alternative to a 'standard' IS or MH approach is the class of methods proposed by Bauwens et al. (2004), adaptive radial-based direction sampling (ARDS) methods, where sampling does not take place in the k-dimensional parameter space directly, but in an $(k-1)$-dimensional subspace of directions. The kth dimension, a distance measure, is drawn from the target distribution itself (conditional on the directions). In this way the

shape of the posterior density is perfectly taken into account along the sampled directions. A disadvantage of the ARDS methods is that evaluations of the target density kernel are required for a grid of points on a line along each sampled direction.

Geweke (1989) proposes a class of split normal or split Student-t distributions: loosely speaking, these are adapted versions of normal or Student-t distributions where (after a normalization of mean and covariance) a different 'standard deviation' is allowed for positive or negative values of each variable. So, for a k-variate split normal density one needs to specify $2k$ 'standard deviations'. Disadvantages of the split normal and split Student-t distribution are that these are necessarily unimodal, and that certain types of shapes involving ridges like the one in the posterior for data of New York in Figure 1.2 can not be well approximated by a split normal or split Student-t distribution.

In chapters 2 and 3 of this thesis, methods are proposed in which neural network functions are used as candidate densities. In fact, the main type of neural network function that is considered in this thesis, is another 'adapted version of the Student-t density': a mixture, or convex combination, of Student-t densities. Like the split Student-t distribution of Geweke (1989), a mixture of Student-t distributions can provide a candidate density with substantial skewness (and high kurtosis). Furthermore, it can deal with multimodality and with non-elliptical shapes due to asymptotes like the ridge around $\pi = 0$ in the posterior of (π, β) in Figure 1.2 for data of New York.

Mixtures of Student-t densities are natural candidate densities for several reasons. First, they can provide an accurate approximation to a wide variety of target densities.[8] Second, this approximation can be constructed by a quick, iterative procedure that is proposed in chapter 2 of this thesis. Third, a mixture of Student-t densities is easy to sample from. Fourth, the Student-t distribution has fatter tails than the normal distribution; especially if one specifies Student-t distributions with few degrees of freedom, the risk is small that the tails of the candidate are thinner than those of the target distribution.

Although the mixture of t densities can be considered as a feed-forward neural network function, one could still ask why the term 'neural network sampling methods' is used throughout this thesis instead of 'mixture sampling methods'. There are basically two reasons for this. First, this stresses that the approximation capabilities are a key property of the mixture of t densities; for it is well-known that several types of neural network functions have a 'universal approximation property'. Second, the mixture of t densities

[8] 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 t densities falls within their framework.

is not the only type of neural network function that is proposed as a candidate density in (chapter 2 of) this thesis. Two other types of neural networks are considered: the exponent of a linear combination of piecewise linear functions, and a linear combination of arctangents (of linear combinations of inputs). However, as is shown in an example in chapter 2, the sampling methods using these two types of networks are considerably slower than the methods based on a mixture of t densities. On the other hand, it should be noted that the latter network, the linear combination of arctangents, has a special property. The arctangent function can be analytically integrated infinitely many times; in chapter 2 formulas are derived for the integrals of the arctangent function. Using these formulas one can analytically evaluate the marginal densities and the moments of a set of random variables of which the density is given by the arctangent-based neural network. Therefore, if one has obtained an arctangent-based network that gives an (almost) perfect approximation to a target density, then one can compute the marginal target densities and moments of the target distribution without requiring any sampling. The formulas for the integrals of the arctangent, that are derived in chapter 2, may also be useful in other applications, for example for the evaluation of derivatives in finance.

In order to illustrate the approximation capabilities of mixtures of t densities, we now consider an example of a highly non-elliptical posterior distribution. Consider the following static 2-regime mixture model in which the variable y_t, the (percentage) quarterly growth rate of the real gross national product (GNP) in the US, has two different mean levels:

$$y_t = \begin{cases} \beta_1 + \varepsilon_t & \text{with probability } p \\ \beta_2 + \varepsilon_t & \text{with probability } 1-p \end{cases}, \qquad t = 1, 2, \ldots, T, \tag{1.4}$$

where ε_t ($t = 1, 2, \ldots, T$) are independent, normally distributed error terms $\varepsilon_t \sim N(0, \sigma^2)$, and where the (non-informative) prior density kernel is specified as $1/\sigma$. For identification it is assumed that $\beta_1 < \beta_2$, so that β_1 and β_2 can be interpreted as the mean growth rates during recessions and expansions, respectively. The parameter p is interpreted as the probability of a recession; the probability of an expansion is $1 - p$.

Figure 1.5 shows the shape of a (conditional) highest posterior density (HPD) credible set for (β_1, β_2, p) where σ^2 is fixed at its posterior mode; this HPD credible set contains points (β_1, β_2, p) for which the posterior density (for quarterly data of the period 1959-2001, shown in Figure 3.2) is higher than at any point outside the set. Like in the instrumental variables regression model, the phenomenon of *local non-identification* can also be found in this model; again this results in a highly non-elliptical posterior distribution, as can be seen in Figure 1.5. If $p = 0$ then β_1 is not identified. Intuitively, this

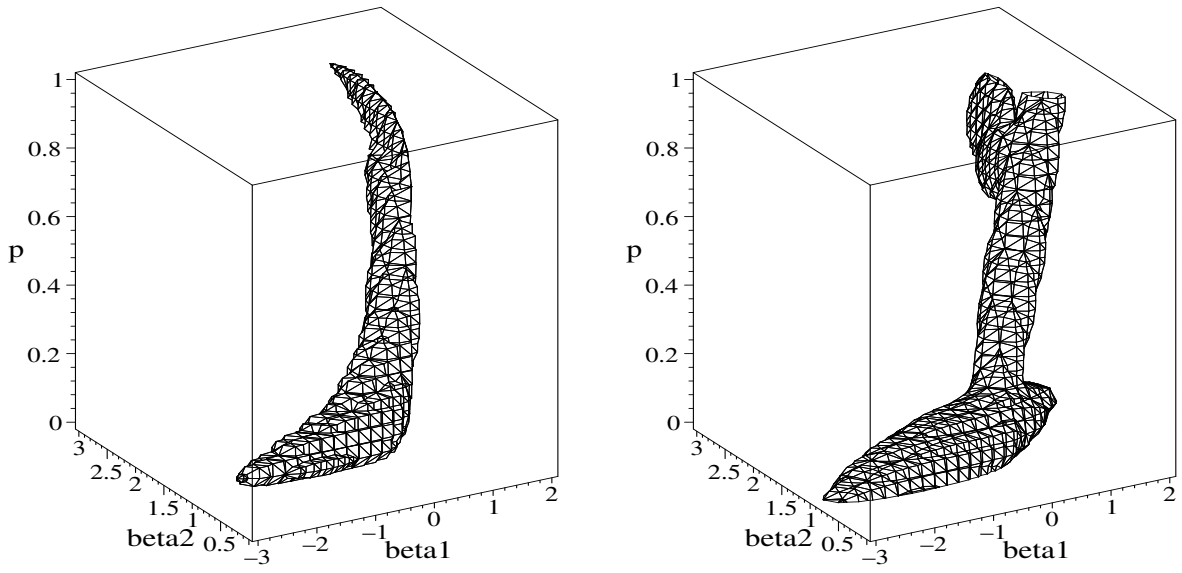

Figure 1.5: Highest posterior density (HPD) credible set (left) and 'highest candidate density' set (right) for a candidate mixture of 5 Student t distributions for parameters (β_1, β_2, p) *in 2-regime mixture model for US real GNP growth rate (conditional on* $\sigma = 0.79$*, the value of* σ *at the posterior mode of* $(\beta_1, \beta_2, \sigma, p)$*)*

reflects that if a recession never occurs, then one cannot identify the mean growth rate during a recession. In a similar fashion, if $p = 1$ then β_2 is not identified. For $p \approx 0$ a wide range of values of β_1 is possible. This reflects that if a recession occurs only rarely, so that one has only few observations on recessions, then there is much uncertainty about the mean growth rate during recessions. In a similar fashion, for $p \approx 1$ a wide range of values of β_2 is possible.

The shape of a 'highest candidate density' set in Figure 1.5, containing points for which the (mixture of t) candidate density is higher than at any point outside the set, now illustrates the ability of mixtures of t densities to provide reasonable approximations to a wide variety of densities. For the shape of the 'highest candidate density' set is similar to that of the HPD credible set. The candidate distribution in Figure 1.5 is a mixture of 5 Student-t distributions. Actually, a mixture of 3 Student-t distributions can already provide a reasonable approximation to the shape of the posterior distribution. This can be seen in Figure 1.6, which illustrates the iterative procedure by which a (mixture of t) candidate distribution is constructed. The procedure starts with a Student-t distribution around the posterior mode. After that, Student-t distributions are iteratively added to the mixture. Each new Student-t distribution is located in an area where the current candidate

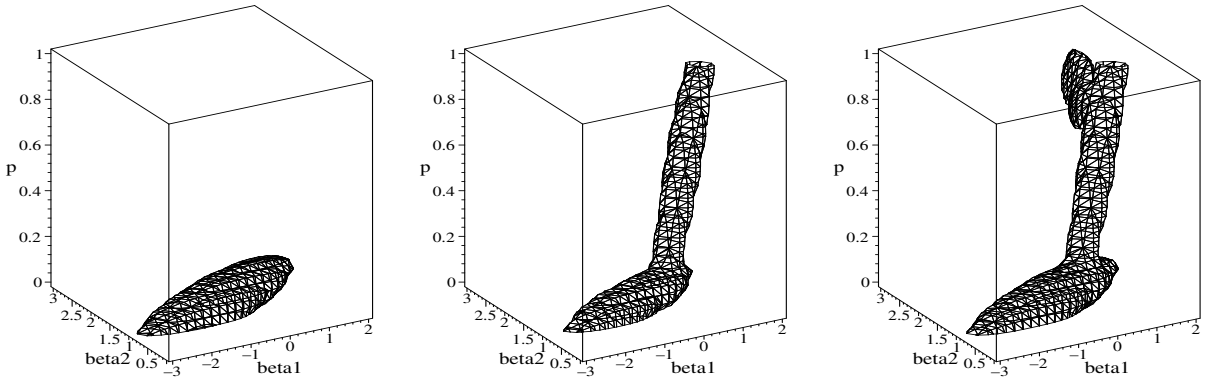

Figure 1.6: 'highest candidate density' sets for a candidate Student-t distribution around the posterior mode (left), a candidate mixture of 2 Student-t distributions (middle), and a candidate mixture of 3 Student-t distributions (right) for parameters (β_1, β_2, p) *in 2-regime mixture model for US real GNP growth rate (conditional on* $\sigma = 0.79$*, the value of* σ *at the posterior mode of* $(\beta_1, \beta_2, \sigma, p)$*)*

density, that has been obtained in the previous step of the algorithm, is too low in the sense that the ratio of the target density and the current candidate density is relatively very high. The iterative method, which starts with a Student-t distribution around the posterior mode, implies that the class of mixtures of t densities is not only useful in extreme cases of highly non-elliptically target (posterior) distributions. If the target is somewhat closer to a normal distribution, then the algorithm may quickly terminate at a mixture of only 2 or 3 Student-t components, which may still provide a substantial improvement over a simple Student-t candidate distribution in IS or the MH procedure. More (technical) details of the iterative construction procedure are given in chapters 2 and 3.

The use of a neural network function, such as a mixture of t densities, as a candidate density can be seen as an investment of computing time, required for the construction of the neural network approximation to the target density, in the quality of the candidate density. This higher quality of the candidate implies a quicker convergence of sampling results and more reliability that no areas with posterior probability mass are 'missed'. This implies that neural network sampling methods are especially useful if one desires very precise estimates of characteristics of the posterior and/or if the posterior is highly non-elliptical. In chapter 3 this result is illustrated, and improvements of (technical details of) the procedure are given which make it even quicker and more reliable.

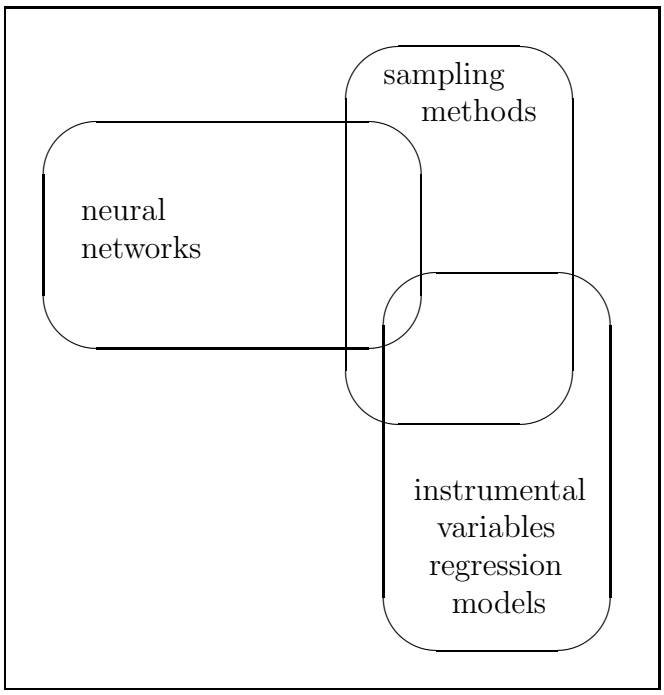

Figure 1.7: Overlap of the three main areas in which the research discussed in this thesis is performed

1.1.3 Structure of the thesis

The research discussed in this thesis concerns areas that can roughly be described by three keywords: neural networks, sampling methods and instrumental variables (IV) regression models. The highly non-elliptical shapes that may occur in posterior distributions in the IV regression model, and the possible usefulness of neural network sampling methods in such cases, explain the overlap of neural networks, sampling methods and IV models as depicted in Figure 1.7.

Chapters 4 and 5 on instrumental variables constitute part II of this thesis. In part II several methods for making inference in instrumental variable regression models are considered and compared in several situations. In part I, consisting of chapters 2 and 3 on neural network sampling methods, the focus is completely different: the (posterior) distribution of interest is treated as given, and the target is to sample from it efficiently, in the sense of yielding reliable results as fast as possible. Hence, the division of this thesis into Part I and Part II.

1.2 Contributions of the thesis

The contributions of this thesis consist of four types of results:

1. *A theoretically perfect candidate density.* A class of neural network functions is introduced that can approximate a wide variety of density functions, and that are easy to sample from (when considered as density functions).

   It should be mentioned that two well-known characteristics of neural networks, that are often seen as disadvantages, are as follows. First, a neural network is a 'black box' in the sense that the working is not completely clear: the values of individual parameters/weights in the network have no straightforward interpretation. Second, neural networks may suffer from 'overfitting', the situation where not only the structural process is captured, but also random noise is 'fitted'. However, in our case these two properties are no disadvantages, as only a good approximation to a target density is required, no interpretation of the neural network parameters is desired; and the data used in the learning process consist of (target density) function evaluations *without random noise*.

2. *An operational procedure for the construction of a useful candidate density, a good approximation to a (possibly highly non-elliptical) target density, in practice.* A quick and reliable algorithm is proposed that constructs a useful candidate distribution, a mixture of t distributions that gives a good approximation to the target distribution for a wide variety of target distributions.

3. *A systematic analysis of the posterior of the parameters in the instrumental variables (IV) regression model under several diffuse prior specifications.* It is shown how the shapes of the posterior distribution depend on instrument strength and the level of endogeneity for several prior specifications. It is illustrated that, although the Jeffreys prior remedies certain peculiar properties that occur under the flat prior, it may still lead to highly non-elliptical posteriors. Further, the hierarchical prior of Chamberlain and Imbens (1996), which also remedies certain peculiar properties that occur under the flat prior, is briefly discussed and compared with the Jeffreys prior. The approach of Chamberlain and Imbens (1996) requires the 'tuning' of a prior variance. The sensitivity of posterior results to the choice of this prior variance clearly suggests that the use of the Jeffreys prior is preferable in most situations.

4. *New empirical results on the (monetary) returns to education:*

   - It is shown that the results on returns to education for the USA of Angrist and Krueger (1991) are completely determined by the region South. If the effect of the return on education is different for the other regions, which can not a priori be ruled out given the large economic differences between these regions, inference using data of the US is not representative for the average returns on education across the US. One should therefore be careful when drawing such conclusions.
   - Bound, Jaeger and Baker (1995) have concluded that the models of Angrist and Krueger (1991) do not give much usable information concerning the causal effect of education on wages. It is shown that this conclusion of Bound, Jaeger and Baker (1995) is too strong, as a model of Angrist and Krueger (1991) (or a slightly modified version) can give usable information on the causal effect of education on income in (regions of) the US.
   - It is shown that quarter of birth is a stronger instrument for education for people with at most 8 or at least 14 years of education than for people with 9-13 years of education. This suggests that quarter of birth does not only affect the number of completed years of schooling for those who leave school as soon as the law allows for it, which is suggested by Angrist and Krueger (1991), as these persons are (mostly) contained in the group with 9-13 years of education. Therefore, if one intends to increase the understanding of the working of the quarter-of-birth instruments, it is a better idea to focus on differences between states in school entry requirements and/or compulsory schooling laws for children of age 5-7 than to concentrate on the differences in compulsory schooling laws for students of age 16-18.

1.3 Outline of the thesis

The outline of this thesis is as follows. In Chapters 2 and 3 it is investigated how a neural network can be used as an importance function in importance sampling or as a candidate density in the Metropolis-Hastings algorithm. Neural networks are natural candidate densities as they can approximate a great variety of functions, and can be specified in such a way that they are easy to sample from, for example as the mixture of t densities mentioned above. In Chapter 2 three types of neural networks and the sampling

methods based on these networks are considered and the performance of these methods is examined in some simple examples. One of the methods that are introduced in Chapter 2 is the AdMit (Adaptive Mixture of t) method, in which a mixture of t distributions approximating the target posterior distribution is first iteratively constructed and then used as the candidate distribution. Chapter 2 is based on Hoogerheide, Kaashoek and Van Dijk (2003a, 2003b, 2004, 2006). In Chapter 3 some improvements of the AdMit method are discussed, and it is applied to the aforementioned posterior distribution in a switching model for the US real GNP. Further, it is illustrated that neural network sampling methods can be especially useful when high precision estimates of posterior characteristics of interest are desired. Chapter 3 is based on Hoogerheide and Van Dijk (2006a).

It should be noted that the applications of neural network sampling methods to IV regression models in Chapters 2 and 3 are mainly for illustrative purposes, that is, to show how the methods can deal with 'extreme' distributions (e.g. with two modes that are far apart). So, the focus is on illustrating the (capabilities of the) sampling methods, not on whether the non-standard characteristics of the posterior distribution could (or should) be circumvented by choosing a different, 'better' model or prior distribution. In other words, in Chapters 2 and 3 the (posterior) distribution of interest is treated as given, and the target is to sample from it efficiently, in the sense of yielding reliable results as fast as possible. In Chapters 4 and 5 the focus is completely different: different methods of performing inference in IV regression models are considered.

In Chapter 4 shapes of posterior distributions in IV regression models are considered for several prior distributions, the flat prior and the Jeffreys prior (for different levels of endogeneity and strength of instruments). Further, a hierarchical prior is briefly discussed and compared with the Jeffreys prior. Chapter 4 is based on Hoogerheide, Kaashoek and Van Dijk (2004, 2006).

In Chapter 5 one of the instrumental variable regression models of Angrist and Krueger (1991) is examined. The differences are considered between the results for data concerning different regions of the US. It also gives a closer examination of some of the assumptions made by Angrist and Krueger (1991). Chapter 5 is based on Hoogerheide, Kleibergen and Van Dijk (2006) and Hoogerheide and Van Dijk (2006b).

In Chapter 6 a summary is given of the main findings in this thesis, and topics for further research are discussed.

Part I

Neural network sampling methods

Chapter 2

Neural networks as candidate densities in importance sampling or the Metropolis-Hastings algorithm

Chapter 2 is based on Hoogerheide, Kaashoek and Van Dijk (2003a, 2003b, 2004, 2006).

2.1 Introduction

Econometric models may be described by the joint probability distribution, known upto a parameter vector θ, of $y = \{y_1, \ldots, y_N\}$, the set of N available observations on the endogenous variable y_i, where y_i may be a vector itself. There are two ways of performing inference on an econometric model: classical and Bayesian inference. In the classical approach the parameters θ are considered as unknown constants that have to be estimated, typically by maximizing the likelihood function $L(\theta) = p(y|\theta)$, the probability density of the data y given a particular value of θ. Bayesian inference proceeds from the likelihood $L(\theta)$ and a prior density $p(\theta)$ reflecting prior beliefs on the parameters before the data set has been observed. So, in the Bayesian approach the parameters θ are considered as random variables of which the prior density $p(\theta)$ is updated by the information contained in the data, incorporated in the likelihood function $L(\theta)$, to obtain the posterior density $p(\theta|y)$. This process is formalized by Bayes' theorem:[1]

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{2.1}$$

[1]Note that this is merely a result of rewriting the identity $p(y)p(\theta|y) = p(\theta)p(y|\theta)$, the two ways of decomposing the joint density $p(y, \theta)$ into a marginal and a conditional.

which can be rewritten as:

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{2.2}$$

where the symbol $\propto$ means "is proportional to", i.e. the left-hand side is equal to the right-hand side times a scaling constant $(1/p(y) = 1/\int p(\theta)p(y|\theta)d\theta)$ that does not depend on the parameters θ.

Typically one is interested in the posterior moments of (and correlations between) the elements of θ, and in the probability that θ belongs to a region D of the parameter space. These characteristics of interest can be expressed as the expectation of a function $g(\theta)$ under the posterior:

$$E[g(\theta)|y] = \int g(\theta)\ p(\theta|y)d\theta = \frac{\int g(\theta)p(\theta)p(y|\theta)d\theta}{\int p(\theta)p(y|\theta)d\theta}. \tag{2.3}$$

From (2.3) it is clear that evaluating integrals is a crucial ingredient in Bayesian inference. The range of models and prior densities for which the integration in (2.3) can be performed analytically is very restricted. In many cases numerical integration methods are required. Basically there are two numerical approaces: deterministic integration and Monte Carlo integration. Deterministic integration consists of evaluating the integrand at a set of many fixed points, and approximating the integral by a weighted average of the function evaluations. Monte Carlo integration is based on the idea that $E[g(\theta)|y]$, the mean of $g(\theta)$ under the posterior, can be approximated by its 'sample counterpart', the sample mean $\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)$, where $\theta_1, \ldots, \theta_n$ are drawn from the posterior distribution.

At a first glance, deterministic integration may always seem a better idea than Monte Carlo integration, as no extra uncertainty (caused by the required random variables) is added to the procedure. However, in deterministic integration the number of required function evaluations increases exponentially with the dimension of the integration problem, which is in our case (2.3) the dimension k of the vector θ. Therefore, deterministic integration approaches like quadrature methods become unworkable if k exceeds, say, three. So, in many cases one has to make use of Monte Carlo integration.[2] However, only for a very limited set of models and prior densities it is possible to directly draw random variables from the posterior distribution. Then one may use indirect sampling algorithms, for example importance sampling or the Metropolis-Hastings algorithm.

Importance sampling, due to Hammersley and Handscomb (1964), was introduced in econometrics and statistics by Kloek and Van Dijk (1978). Roughly speaking, importance

[2]In models with latent variables, the likelihood is itself an integral; if this integral can not be evaluated analytically, one may need Monte Carlo integration in classical inference in these models.

sampling consists of drawing a set $\theta_1, \ldots, \theta_n$ from a candidate density $q(\theta)$, also known as the importance function, and approximating $E[g(\theta)|y]$ by the weighted average of $g(\theta_1), \ldots, g(\theta_n)$ with weights (adding to one and) proportional to $w(\theta_i)$ with $w(\theta) \equiv p(\theta|y)/q(\theta)$, where $p(\theta|y)$ may be the kernel $L(\theta)p(\theta)$ of the posterior density; it does not have to be the posterior density itself:

$$\hat{g}_{IS} = \frac{\sum_{i=1}^{n} g(\theta_i)w(\theta_i)}{\sum_{i=1}^{n} w(\theta_i)}. \tag{2.4}$$

Importance sampling is based on the relationship:

$$E[g(\theta)|y] = \frac{\int g(\theta)p(\theta|y)d\theta}{\int p(\theta|y)d\theta} = \frac{\int g(\theta)w(\theta)q(\theta)d\theta}{\int w(\theta)q(\theta)d\theta} = \frac{E[g(\tilde{\theta})w(\tilde{\theta})]}{E[w(\tilde{\theta})]}, \tag{2.5}$$

where $\tilde{\theta}$ is a random variable with density $q(.)$, the candidate density. Under certain conditions $\hat{g}_{IS}$ in (2.4) is a consistent estimator of $E[g(\theta)|y]$ in (2.5).

The Metropolis-Hastings (MH) algorithm is a Markov chain Monte Carlo (MCMC) approach that has been introduced by Metropolis et al. (1953) and generalized by Hastings (1970). Markov chain Monte Carlo methods construct a Markov chain converging to a target distribution, in our case the posterior distribution of interest. After a burn-in period, which is required to make the influence of initial values negligible, draws from the Markov chain are considered as (correlated) draws from the target distribution itself.

In the MH algorithm a Markov chain of length n is constructed by the following procedure. First, one chooses a feasible initial state θ_0. Then one repeats the following steps n times (for $i = 1, \ldots, n$). A candidate value $\tilde{\theta}_i$ is drawn from the candidate transition density $q(\theta_{i-1}, .) = q(.|\theta_{i-1})$, and a random variable U is drawn from the uniform distribution $U(0, 1)$. Then the acceptance probability (or transition probability)

$$\alpha(\theta_{i-1}, \tilde{\theta}_i) \equiv \min \left\{ \frac{p(\tilde{\theta}_i|y)q(\tilde{\theta}_i, \theta_{i-1})}{p(\theta_{i-1}|y)q(\theta_{i-1}, \tilde{\theta}_i)}, 1 \right\} \tag{2.6}$$

is computed. If $U < \alpha(\theta_{i-1}, \tilde{\theta}_i)$ then the transition to the candidate value is accepted: $\theta_i = \tilde{\theta}_i$. Otherwise the transition is rejected, and the next state is again θ_{i-1}, that is $\theta_i = \theta_{i-1}$.

The candidate transition density q can be specified in several ways. For two common specifications the MH algorithm yields an independence chain or a random walk chain. In the independence chain MH algorithm the candidate $\tilde{\theta}_i$ is drawn independently from the current state θ_{i-1}; the candidate density is the same for each $i = 1, \ldots, n$:

$q(\theta_{i-1}, \tilde{\theta}_i) = q(\tilde{\theta}_i)$. In this case the acceptance probability is given by $\alpha(\theta_{i-1}, \tilde{\theta}_i) = \min\{w(\tilde{\theta}_i)/w(\theta_{i-1}), 1\}$, where the occurrence of the importance weights $w(\theta) \equiv p(\theta|y)/q(\theta)$ shows a link between the independence chain MH algorithm and importance sampling.

In the random walk MH algorithm the candidate transition step $\tilde{\theta}_i - \theta_{i-1}$ is drawn instead of the candidate state $\tilde{\theta}_i$: $q(\theta_{i-1}, \tilde{\theta}_i) = q(\tilde{\theta}_i - \theta_{i-1})$. A common choice for the distribution of the candidate step is a normal or Student-t distribution with mode 0. For a more elaborate overview of numerical integration methods, both deterministic and Monte Carlo integration methods, the reader is referred to Van Oest (2005). An overview of Monte Carlo integration methods can also be found in Hoogerheide, Van Dijk and Van Oest (2006). Extensive discussions on solely deterministic integration methods are given by Stoer and Bulirsch (1993) and Cheney and Kincaid (1994).

For both importance sampling and the independence chain MH algorithm it holds that the candidate density should be 'close' to the target density, and it is especially important that the tails of the candidate should not be thinner than those of the target.

For importance sampling, Geweke (1989) shows that the optimal importance density for estimating $E[g(\theta)|y]$ has kernel $|g(\theta) - E[g(\theta)|y]|\, p(\theta|y)$, where optimal means having the minimal (asymptotically valid) variance. However, this result is of limited practical relevance for three reasons. First, one would need to obtain a preliminary estimate of $E[g(\theta)|y]$ using a different, less efficient method. Second, a different importance function (and a corresponding set of draws) would be required for each different characteristic of interest $E[g(\theta)|y]$. Finally, a method for sampling from this optimal importance density would have to be devised. On the other hand, the expression for the optimal importance density does imply that an importance density with tails thicker than the posterior density might be more efficient than the posterior density itself, as the tails of the density with kernel $|g(\theta) - E[g(\theta)|y]|\, p(\theta|y)$ decay slower than the tails of $p(\theta|y)$.

Although the optimal importance density depends on both the posterior density and the characteristic of interest, it is impractical to choose a different importance density for each characteristic of interest; Geweke(1989) suggests that when considering a class of possible importance densities $q(\tilde{\theta})$ it is a reasonable objective to choose the one that minimizes $E[w(\tilde{\theta})^2] \propto \int p(\tilde{\theta}|y)^2/q(\tilde{\theta})d\tilde{\theta}$.

In a 'standard' case of importance sampling or the independence chain MH algorithm, the candidate is unimodal; a common choice is a normal or Student-t distribution. If the target (posterior) distribution is bimodal then a second mode may be completely missed

in the MH approach and some draws may have huge weights in importance sampling. So, for certain non-elliptical target distributions the convergence behavior of these 'standard' Monte Carlo integration methods is rather uncertain. In these cases one needs to look for either different Monte Carlo integration methods or a more appropriate candidate density than a normal or Student-t distribution.

One such alternative to a 'standard' importance sampling or MH approach is the class of methods proposed by Bauwens et al. (2004), adaptive radial-based direction sampling (ARDS) methods, where sampling does not take place in the k-dimensional parameter space directly, but in an $(k-1)$-dimensional subspace of directions. The kth dimension, a distance measure, is drawn from the target distribution itself (conditional on the directions). In this way the shape of the posterior density is perfectly taken into account along the sampled directions.

Geweke (1989) proposes a class of split normal or split Student-t distributions: loosely speaking, these are adapted versions of normal or Student-t distributions where (after a normalization of mean and covariance) a different 'standard deviation' is allowed for positive or negative values of each variable. So, for a k-variate split normal density one needs to choose $2k$ 'standard deviations' for the k elements of θ.

In this chapter we propose methods in which neural network functions are used as candidate densities. In fact, one of the proposed types of neural networks is another 'adapted version of the Student-t density': a mixture of Student-t densities. Like the split Student-t distribution of Geweke (1989) a mixture of Student-t distributions can provide a candidate density with skewness (and high kurtosis); furthermore, it can deal with multimodality. This is exactly why neural networks are natural candidate densities; they have a 'universal approximation' property, which means that neural network functions can approximate a great variety of 'non-standard' density functions. Moreover, neural networks can be specified in such a way that they are easy to sample from.

Section 2.2 gives a short overview of neural networks in general. Section 2.3 describes three types of neural networks that are easy to sample from, and ways to construct neural network approximations to a target density. In section 2.4 the performance of these neural networks is compared in a simple example. In section 2.5 some neural network sampling methods are applied to an illustrative example, in which their performance is compared with other algorithms such as importance sampling using a Student-t candidate, a random walk Metropolis-Hastings algorithm and Gibbs sampling. Section 2.6 contains concluding remarks.

2.2 Neural networks

2.2.1 What is a neural network?

An obvious, simple question one could ask is "what is a neural network?". However, although this is a clear, simple question, it is not so easy to give an answer about which every researcher would agree, as there exists no precise definition that is generally accepted among researchers. A reason for this is that the field of neural networks has evolved independently in and among many sciences such as neurobiology, mathematics, computer science and psychology.[3]

First of all, it should be noted that neural networks are either biological or artificial neural networks; a biological neural network is a system of physically interconnected neurons such as the (human) brain, while an artificial neural network (ANN) is a mathematical model or artificially created (computer) system that is merely inspired by the working of the human brain. In the human brain, a neuron (i.e. a cell in the nervous system) collects signals from others through a system of fine structures called 'dendrites'. When a neuron receives sufficiently large input, it sends a spike of activity down a long, thin cable-like part called the 'axon'. Neurons have only one axon, but this axon may undergo extensive branching, enabling communication with many target cells. At the end of each branch a structure called 'synapse' transforms the signal from the axon into effects that may cause activity in a connected neuron. Whether or not a neuron sends a signal to connected neurons depends on whether the *sum* of these effects coming from the synapse goes above a certain threshold (which differs between neurons); this principle is called 'neural summation'. Learning occurs by changing the effectiveness of the synapses so that the influence of neurons on each other changes.

Biological neurons communicate with multiple types of signals, both chemical and electrical; moreover, their electrical output usually consists of complex sequences of spikes instead of a simple potential. It should also be noted that the knowledge about the working of neurons is still incomplete. Therefore artificial neural networks are necessarily (gross) simplifications of biological neural networks.

[3]One consequence is that entering 'define:Neural network' at `www.google.com` yields 20 neural network definitions; between some of these definitions there are huge differences: for example, in different definitions a neural network is said to be an 'interconnected group of neurons','method for optimizing', 'member of a class of software', or 'modeling technique'. One of these definitions comes from the Wikipedia site `en.wikipedia.org/wiki/Neural_network` where also much information about several specific types of neural networks can be found.

In this thesis only artificial neural networks will be considered. However, also for artificial neural networks there is no generally accepted definition, as will be discussed below. In the literature the 'artificial' is often dropped as it is mostly perfectly clear from the context whether or not one refers to artificial neural networks. Throughout the rest of this thesis this convention is followed: the term 'neural network' is used, and always denotes an artificial neural network.

Stergiou and Siganos (1996) give a definition of an (artificial) neural network as an information processing paradigm that is inspired by the way biological nervous systems (such as the brain) process information, in the sense that it consists of a number of interconnected processing elements working together to solve specific problems, and in the sense that it 'learns by example': each neural network is configured for a specific application through a 'learning process'.

The Defense Advanced Research Projects Agency (DARPA) Neural Network Study (1988) gives the following definition of a neural network: "a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strength, and the processing performed at computing elements or nodes."

The first definition nicely summarizes some of the typical neural network characteristics in a sentence. The latter definition tells some of the 'building blocks' that make up a neural network. The latter approach is extended by Fiesler (1994) who elaborately discusses the 'ingredients' that generally make up a neural network, resulting in a broad neural network definition, broad enough to encompass both simple biological neural networks, and essentially all artificial neural networks. Fiesler gives a definition of a neural network as a set of 4 characteristics:

1. **topology.** The topology concerns the *frame(work)* and the *interconnection structure* of the neural network. The frame(work) is defined by the number of clusters (i.e. groups of neurons, the elements of the network, into which the network may be divided), and the number of neurons in each cluster. The interconnection structure is the set of relations between neurons: this does not only define which neurons are connected, but also whether connections are symmetric (i.e. bidirectional, having the same weight/interconnection strenght in either direction) or asymmetric (i.e. unidirectional, where the weight is only used for propagation in one direction), and whether the clusters of the network are 'slabs' (clusters that have a similar function or hierarchical level) or 'layers' (clusters that have a natural ordering).

Another neural network characteristic that is part of the interconnection structure is the 'order' of connections. A high(er) order connection combines inputs of several neurons by another way than (weighted) summation, usually by multiplication; the number of inputs that is combined in a non-linear way is the 'order' of the connection. The order of a neural network is the highest order of the connections in it. Most (traditional) neural networks are first order neural networks, i.e. neural networks in which any non-linearity only occurs by means of application of a non-linear (activation) function to the weighted sum of inputs in the neuron.[4] Moreover, it should be noted that because of the principle of 'neural summation' in biological neural networks, by which artificial neural networks are inspired, one could argue that a high(er) order network does not really belong to the class of neural networks.

2. **constraints.** The constraints define the value ranges for the weights of the connections, and the thresholds and activation values of the neurons: for example, one may want to restrict (certain) weights to lie within a certain interval or set of integers.

3. **initial state.** The initial state is the set of initial values for weights, thresholds, and (input neuron) activations. These are the values before the 'learning process' has started.

4. **transition functions.** There are four types of transition functions:

   (i) *activation functions* (also known as neuron functions or transfer functions), which specify the output of a neuron given its inputs;

   (ii) *learning rules*, defining how weights (and thresholds) will be updated during the learning process;

   (iii) *clamping functions*, determining if and when certain neurons retain their present activation value during the learning process;

   (iv) *ontogenic functions*, specifying changes in the neural network topology.

   The types of functions (i) and (ii) occur in almost all neural networks, while the types (iii) and (iv) can be only found in specific networks. The networks used in this thesis contain functions of the types (i), (ii) and (iv).

[4] In a well-known popular class of neural networks, radial basis function (RBF) neural networks, the output of certain neurons is a (non-linear) function of the sum of squared deviations of inputs from weights. However, considering these as neural networks in which certain neurons only get one input signal and yield a squared deviation as output, these are also first order neural networks.

This definition encompasses several types of neural networks. One extreme type is the 'plenary neural network', a neural network in which all neurons are connected with each other. The plenary neural network is a special (extreme) case of a 'feedback neural network', a network in which signals travel in different directions, so that its 'state' continuously changes until it reaches an equilibrium point. The opposite of a feedback network is the 'feed-forward neural network', in which signals are allowed to travel in only one way, from input(s) to output(s): there is no feedback in the sense of 'loops'. In feed-forward networks outputs can be calculated as explicit functions of inputs and weights; in fact, feed-forward networks can be considered as a general framework for representing (non-linear) functional mappings between a set of input variables and a set of output variables.

In this thesis we only make use of certain layered first order feed-forward (artificial) neural networks: feed-forward (artificial) neural networks in which neurons are grouped in three or four layers of neurons[5] that are naturally ordered from input to output, and in which any non-linearity stems from activation functions that are applied to weighted sums of inputs.

Confining ourselves to (layered) feed-forward (artificial) neural networks makes it easier to find an agreed definition. Crespin (1995) gives a definition in terms of three elementary concepts: a feed-forward neural network is *parametric composition* of layers, where a layer is a *parametric product* of neurons, where a neuron is a *parametric map*. These three concepts are defined as follows:

- A *parametric map* with domain X, range Y and with parameters in parameter space W is a function $f : W \times X \to Y$.

  Note that in the first order neural network, the most common network, the function f should satisfy $f(w, x) = \tilde{f}(\sum_i w_i x_i)$ with parameters $w \in W$, input $x \in X$ and the summation over all elements in w, x for some function $\tilde{f} : \mathbb{R} \to Y$.

- The *parametric product* of n parametric maps $f_j : W_j \times X \to Y_j$ $(j = 1, \ldots, n)$, all with common input $x \in X$, is the map $\hat{\Pi}_{j=1}^{n} f_j : W_1 \times \cdots \times W_n \times X \to Y_1 \times \cdots \times Y_n$ defined as $\hat{\Pi}_{j=1}^{n} f_j(w_1, w_2, \ldots, w_n; x) = (f_1(w_1, x), \ldots, f_n(w_n, x))$. Note that a parametric product of parametric maps is a parametric map itself.

[5]We follow the convention of counting layers of neurons. There is also another convention, used in for example Bishop (1995), in which layers of (adaptive) weights instead of neurons are counted; in that sense we use networks of two or three layers.

- The *parametric composition* of p parametric maps $f^k : W^k \times X^k \to X^{k+1}$ $(k = 1, \ldots, p)$ is the map $\widehat{\bigcirc}_{k=1}^{p} f^k$ defined recursively by $(\widehat{\bigcirc}_{k=1}^{p} f^k)(w^1, \ldots, w^p, x^1) = f^p(w^p, (\widehat{\bigcirc}_{k=1}^{p-1} f^k)(w^1, \ldots, w^{p-1}, x^1))$.

We conclude that for a (layered) feed-forward artificial neural network it is possible to give a much simpler definition.

Although the definition in terms of 'building blocks' of Fiesler (1994) is well formulated and rather general, it may be intuitively less appealing than the first of the definitions at the beginning of this section. Likewise, a biologist may not be satisfied with an overview of body parts as an answer to the question what a monkey is. We end this subsection with a definition that is at least as general as Fiesler's, and that tries to say what a neural network is instead of just summing up 'ingredients'. A neural network is a mathematical model or artificially created (computer) system (in case of an artificial NN), or group of physical neurons (in case of a biological NN) that has the following characteristics:

- A neural network consists of *(many) interconnected elements*.
- A neural network has to be trained for a specific application: it *'learns by example'*. Except for the most trivial networks, one can not a priori choose plausible values for its connection strengths (or weights) and thresholds based on for example economic theory, unlike coefficients in certain linear models which may be expected to equal a certain value from theory.
- A neural network *stores its processing capability in the connection strengths between its elements*.
- In each neuron the *sum* of the effects caused by signals from other neurons determines whether the neuron sends a signal itself (and also what kind of signal it sends). There are researchers who use neural networks with high(er) order connections, in which (functions of) inputs may be combined in a different way than summation. However, these networks are exceptions, and one could argue that these are not truly neural networks. Therefore, we conclude that the *additive structure* is a typical property of a neural network .

The fact that neural networks have been used (and are still used) in several different sciences has not only caused the existence of many different definitions; there also exist several names for some of the 'ingredients' in neural networks. Table 2.1 gives the names

for 'building blocks' in neural networks that will be used throughout (the rest of) this thesis and some synonyms.

Table 2.1: Terms used throughout (the rest of) this thesis and some synonyms that are used in neural network literature

name used throughout the rest of this thesis	synonyms
activation function	neuron function, transfer function
input	independent variable, point in domain
layer	cluster*
neuron	cell, node, processing unit, processor, unit
output	dependent variable, image point
weight	parameter, control, connection strength

*actually the term 'cluster' is more general, as clusters are either slabs or layers, where the latter have some (natural) ordering.

2.2.2 Why and when can a neural network be useful?

After trying to answer the question "what is a neural network?" in the previous subsection, this subsection aims to answer the question "why and when can a neural network be useful?". In general, because they are inspired by the human brain, neural network based computer programs are able to solve some problems that people are good at solving, but at which other computer programs perform poorly. One such application is given by pattern recognition, for example handwritten word recognition, recognition of speakers in communications, facial recognition, or 3-dimensional object recognition (e.g. interpreting sonar traces).

The strength of neural networks is that the interaction of the (large number of) elements of the network can generate very complex maps from inputs to outputs. As a result, neural networks are able to approximate a great variety of functions. This 'universal approximation property' is the reason why neural networks are used in this thesis, as we want to find candidate density functions that in some sense approximate target density functions. In this thesis several neural network specifications will be used. These specifications result in different capabilities and the proofs of these capabilities have been

given in different publications. Therefore, references for these proofs will be given after the specific neural networks have been described.

Like all models, the class of neural networks also has some drawbacks; we mention three well-known disadvantages of neural networks that are often discussed in literature. First, a neural network is a 'black box' in the sense that the working is not completely clear: the values of individual parameters have no straightforward interpretation and, even in the case of simple, monotonous activation functions, neural networks can generate very complex maps from inputs to outputs. Second, neural networks may suffer from 'overfitting', the situation where not only the structural process is captured, but also random noise is 'fitted'. This problem may also be present in linear models, but in neural networks preventing it is more difficult, as one may increase the number of parameters by adding (hidden) neurons that do not correspond with inputs or outputs, without adding extra data on inputs or outputs. Third, the 'learning process' of the neural network may require large computational power.

In the neural network applications in this thesis the disadvantages have the following implications. The first drawback is not so much of a disadvantage in our case; only a good approximation to a target density is required, no interpretation of the neural network parameters. However, a neural network's ability to generate complex patterns implies that one should check the network's behaviour for other input values than those for which it is 'trained', as it may be dangerous to simply extrapolate (or even interpolate) neural network behaviour. The second drawback, the danger of 'overfitting', implies no problems at all in our application, as the data used in the learning process consist of (target density) function evaluations *without random noise*. The third disadvantage has obviously become a smaller problem over time, as the computer capabilities have enormously increased. However, also other computer programs profit from the higher computing speed, of course; so it is of interest to compare the speed of neural network based programs with other methods, as will be done in the sequel of this chapter.

Finally, as mentioned by Stergiou and Siganos (1996), a number of scientists argue that 'consciousness' is a mechanical property and that 'conscious' artificial neural networks are a realistic possibility. If this would be theoretically possible, increased insight in the working of the human brain and better computational facilities may enable researchers to accomplish this some day, which is arguably the most exciting application of an artificial neural network one could think of.

2.3 Neural networks that are easy to sample from

Consider a certain distribution, for example a posterior distribution, with density kernel $\tilde{p}(\theta)$ with $\theta \in \mathbb{R}^k$. Notice that in section 2.1 the notation $p(\theta|y)$ is used to indicate a posterior density (kernel) in order to stress its meaning and the difference with the prior density and likelihood function; here we denote the target density kernel by $\tilde{p}(\theta)$ as only one (target) density function of θ appears in this section, and also because the exposed methods can be applied to other distributions than posteriors as well. Suppose the aim is to investigate some of the characteristics of $\tilde{p}(\theta)$, for example the mean and/or covariance matrix of a random vector $\theta \sim \tilde{p}(\theta)$. The approach followed in this thesis consists of the following steps:

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

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

$$nn(\theta) = eG_2\left(CG_1(A\theta + b) + d\right) + f, \qquad \theta \in \mathbb{R}^k, \tag{2.7}$$

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

$$G_1(v) = (g_1(v_1), \cdots, g_1(v_{H_1}))', \quad G_2(z) = (g_2(z_1), \cdots, g_2(z_{H_2}))', \quad v \in \mathbb{R}^{H_1}, \;\; z \in \mathbb{R}^{H_2} \tag{2.8}$$

where $g_1 : \mathbb{R} \to \mathbb{R}$ and $g_2 : \mathbb{R} \to \mathbb{R}$ are the activation functions. The network with functional form in (2.7) is a first order network, i.e. a network in which each (non-linear) activation function is only applied to a linear combination of the inputs of a neuron.

Figure 2.1 shows (for the case with $k = 2$, $H_1 = 2$, $H_2 = 2$) the *network diagram* representing this 4-layer feed-forward neural network: there is a 1-to-1 correspondence between components of the function (2.7) and elements of the diagram. The circles in the input layer represent the k input variables and a 'bias' term which is simply an extra input variable whose value is permanently set at 1; this bias implies that the capabilities

of the neural network does not depend on the mean of the input variables. There is also a bias term in each hidden layer. Each black (filled) circle in the hidden layers represents a hidden neuron which gives as output a (possibly non-linear) activation function evaluated at the weighted sum of its inputs. For the H_1 (H_2) neurons in the first (second) hidden layer this activation function is given by g_1 (g_2). The circle in the output layer obviously represents the output $nn(\theta)$ of the neural network, the function (2.7) evaluated at θ.

The arrows show which neurons influence which other neurons and the coefficients at the arrows indicate the magnitude of this influence (i.e. the weight in the weighted sum of inputs). This network has connections from every neuron in one layer to every neuron in the next layer, but no other connections are allowed. This is a popular way of specifying a feed-forward network, as it is relatively easy to analyze and implement.

In this thesis the number of layers of neurons is counted, and non-linear transformations are only allowed in hidden layer neurons. However, it should be noted that there is also another convention, used in for example Bishop (1995), in which non-linear transformations are also allowed in the output layer, and layers of (adaptive) weights instead of neurons are counted; in that sense the network in Figure 2.1 is a 3-layer network. It can also be denoted as a double hidden-layer network to prevent misunderstandings.

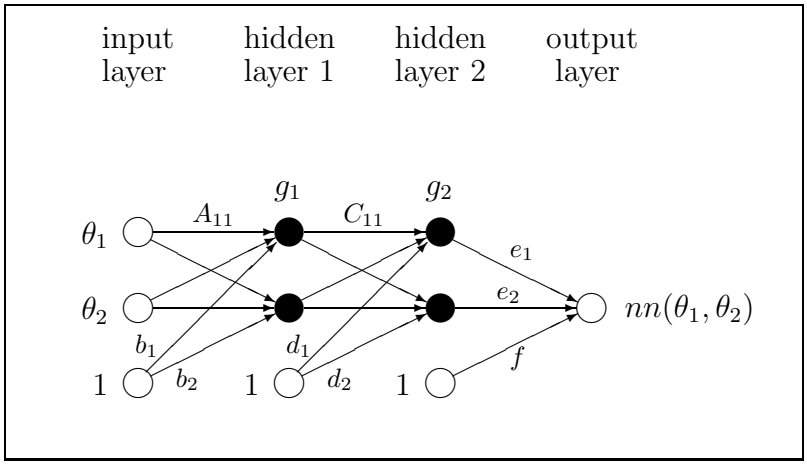

Figure 2.1: Network diagram corresponding to the four-layer neural network function (2.7) in case of $k = 2$ inputs, $H_1 = 2$ neurons in the first hidden layer, $H_2 = 2$ neurons in the second hidden layer. Black (filled) circles represent hidden layer neurons in which non-linear processing takes place. White circles represent input neurons, output neurons or 'bias terms'.

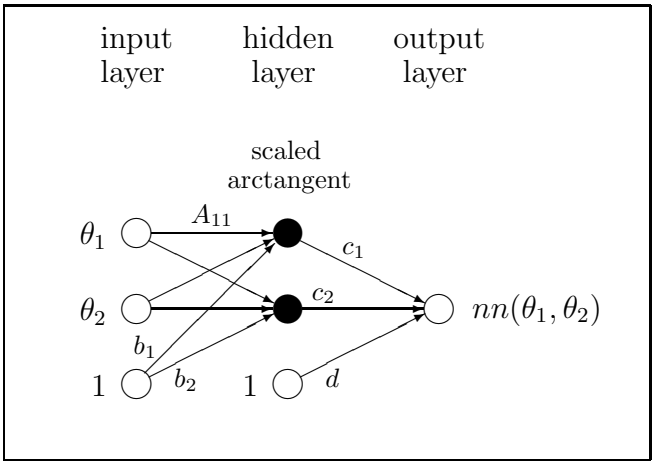

Figure 2.2: Network diagram corresponding to the Type 1 neural network in case of $k = 2$ inputs and $H_1 = 2$ neurons in the hidden layer.

The following three specifications of (2.7) allow for easy sampling (when this neural network function is considered as a density kernel):

Type 1 neural network: A standard three-layer feed-forward neural network (in the notation of (2.7): $H_2 = 1$, $e = 1$, $f = 0$ and g_2 is the identity $g_2(x) = x$, $x \in \mathbb{R}$). As activation function g_1 in (2.8) we take the scaled arctangent function:

$$g_1(x) = \frac{1}{\pi} \arctan(x) + \frac{1}{2}, \qquad x \in \mathbb{R}. \tag{2.9}$$

The reason for choosing the arctangent function is that it can be analytically integrated infinitely many times. We show in subsection 2.3.2, that this property makes the neural network, in the role of a density kernel on a bounded region, easy to sample from. The scaling is merely done because it is common practice to use activation functions that take values in the unit interval. Figure 2.2 shows (for the case with $k = 2$, $H_1 = 2$) the network diagram representing the Type 1 neural network. Such a multi-layered feed-forward network having either sigmoidal or threshold activation functions is known as a *multi-layer perceptron* (MLP); multi-layer perceptrons are generalizations of the *perceptron* of Rosenblatt (1962), a 3-layer network in which only the second layer of weights was updated during the learning process, that was mostly used for the classification ('perception') of binary images of characters.

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

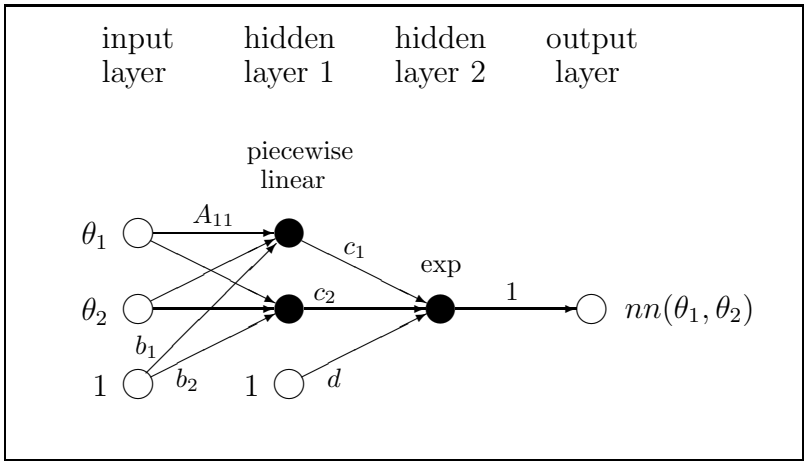

Figure 2.3: Network diagram corresponding to the Type 2 neural network in case of $k = 2$ inputs and $H_1 = 2$ neurons in the hidden layer.

activation function g_1 in (2.8) equal to the following piecewise-linear function *plin*:

$$plin(x) = \begin{cases} 0 & x < -1/2 \\ x + 1/2 & -1/2 \leq x \leq 1/2 \\ 1 & x > 1/2 \end{cases}, \qquad x \in \mathbb{R}. \tag{2.10}$$

We show in subsection 2.3.2 that these activation functions make Gibbs sampling (see Geman and Geman (1984)) possible. To allow for easy sampling it is sufficient to specify a function g_2 which is positive-valued and has an analytical expression for its primitive that is analytically invertible; see subsection 2.3.2. Another example of such a function is the logistic function. Figure 2.3 shows (for the case with $k = 2$, $H_1 = 2$) the network diagram representing the Type 2 neural network.

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

$$nn(\theta) = \sum_{h=1}^{H} p_h \, t(\theta|\mu_h, \Sigma_h, \nu), \tag{2.11}$$

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 t components and where $t(\theta|\mu_h, \Sigma_h, \nu)$ is a k-variate t density with mode vector μ_h, scaling matrix Σ_h, and ν degrees of freedom:

$$t(\theta|\mu_h, \Sigma_h, \nu) = \frac{\Gamma((\nu + k)/2)}{\Gamma(\nu/2)(\pi\nu)^{k/2}} \, |\Sigma_h|^{-1/2} \left(1 + \frac{(\theta - \mu_h)' \Sigma_h^{-1} (\theta - \mu_h)}{\nu}\right)^{-(\nu+k)/2}. \tag{2.12}$$

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

that this mixture of t densities is a four-layer feed-forward neural network (with parameter restrictions) in which we have, in the notation of (2.7), $H_2 = H$ (the number of t densities), $H_1 = Hk$, activation functions

$$g_1(x) = x^2 \quad \text{and} \quad g_2(x) = x^{-(\nu+k)/2}, \qquad x \in \mathbb{R},$$

with weights $f = 0$ and

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

$$e_h = p_h \, |\Sigma_h|^{-1/2} \, c_{\nu,k} \quad (h = 1, \ldots, H) \qquad \text{with} \quad c_{\nu,k} \equiv \frac{\Gamma((\nu+k)/2)}{\Gamma(\nu/2)(\pi\nu)^{k/2}},$$

where ι_m denotes a $m \times 1$ vector of ones. Notice that $(\theta - \mu_h)'\Sigma_h^{-1}(\theta - \mu_h)$ is the sum of the squared elements of $\Sigma_h^{-1/2}(\theta - \mu_h)$. Figure 2.4 shows (for the case with $k = 2$, $H = 2$ so that $H_1 = Hk = 4$, $H_2 = H = 2$) the network diagram representing the Type 3 neural network.

The Type 3 network can also be interpreted as a 3-layer network of *order* h, containing H hidden neurons transforming θ into $t(\theta|\mu_h, \Sigma_h, \nu)$ for $h = 1, \ldots, H$.[6] Figure 2.5 shows (for the case with $k = 2$, $H = 2$) the corresponding network diagram. The Type 3 network is a (3-layer) *Radial Basis Function* (RBF) neural network, a network in which the activation of hidden neurons is determined by the *distance* between the input vector and a certain vector of weights, possibly allowing for a pre-multiplication of the input vector taking care of the covariance between the inputs. The output of a RBF network is a linear combination of its basis functions; the Type 3 network has basis functions $t(\theta|\mu_h, \Sigma_h, \nu)$. Note that, similar to the Type 3 network, all RBF networks can be considered as first order networks (with restrictions on certain weights).

During the past 50 years many results on the approximation capabilities of neural networks have been published. The theorem of Kolmogorov (1957) is often seen as a basis for the (approximation) capabilities of first order 4-layer feed-forward networks.

[6]Recall from the introduction to this chapter that the *order* of a neural network is the highest order of the connections in the network; for a high order (i.e. an order higher than 1), the order of a connection is the number of inputs combined in a different way than (applying an activation function to) a weighted sum.

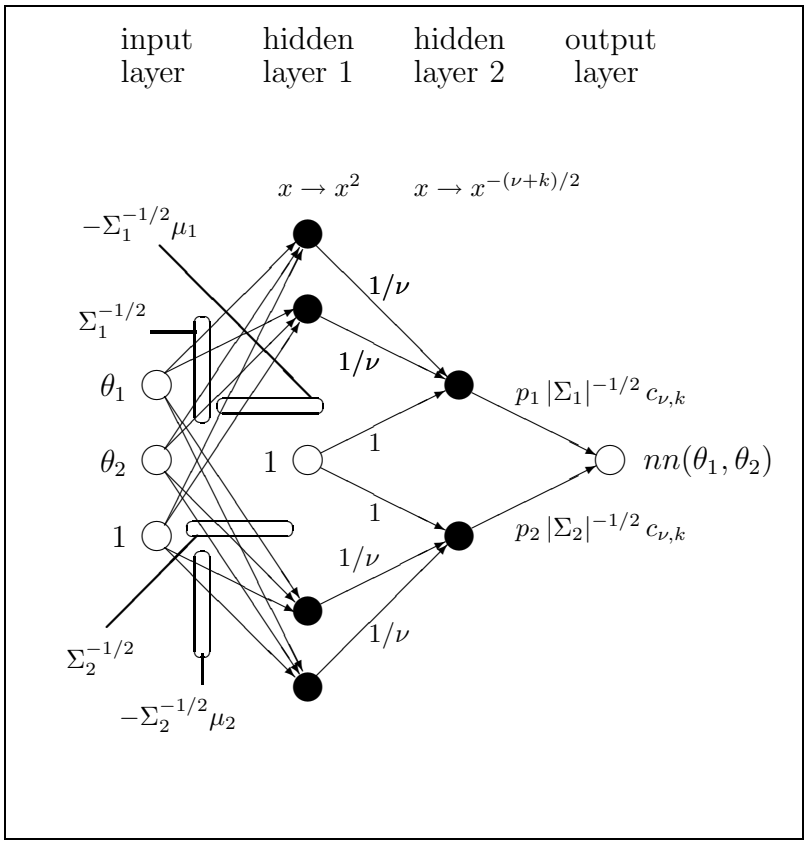

Figure 2.4: Network diagram corresponding to the Type 3 neural network in case of $k = 2$ *inputs and* $H = 2$ *mixture components, interpreted as a first order 4-layer network with* $H_1 = Hk = 4$ *and* $H_2 = H = 2$ *neurons in the first and second hidden layer, respectively. The weights between input layer and hidden layer 1 are elements of either* $\Sigma_h^{-1/2}$ *or* $-\Sigma_h^{-1/2}\mu_h$ *(*$h = 1, 2$*);* $c_{\nu,k}$ *is given by* $c_{\nu,k} \equiv \Gamma((\nu + k)/2)/\{\Gamma(\nu/2)(\pi\nu)^{k/2}\}$.

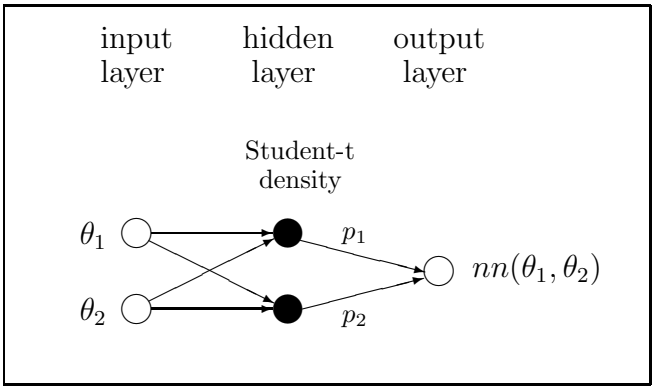

Figure 2.5: Network diagram corresponding to the Type 3 neural network in case of $k = 2$ inputs and $H = 2$ mixture components, interpreted as a high-order 3-layer Radial Basis Function (RBF) network

Kolmogorov's theorem says that (for a closed and bounded input domain) any continuous mapping $\tilde{f}$ from k input variables x_i to an output variable $y = \tilde{f}(x_1, \ldots, x_k)$ can be represented *exactly* by a certain 4-layer network with $k(2k+1)$ and $2k+1$ neurons in the first and second hidden layer, respectively; the functional form of this network is:

$$y = nn(x_1, \ldots, x_k) = \sum_{j=1}^{2k+1} g\left(\sum_{i=1}^{k} \lambda_i \, h_j(x_i)\right)$$

with constants $0 < \lambda_i < 1$ $(i = 1, \ldots, k)$, continuous function g that depends on the function $\tilde{f}$, and functions h_j $(j = 1, \ldots, 2k+1)$ that do not depend on $\tilde{f}$. Although Kolmogorov's theorem gives a nice illustration of the representation/approximation power of neural networks, it is hardly relevant in practical situations because of the following limitations. First, Kolmogorov (1957) did not describe a method to construct the functions g and h_j; in fact, no examples of functions g and h_j are known. Second, in practice one mostly uses fixed activation functions and only adjusts the weights and the number of hidden neurons. Intuitively speaking, it is easier to search through spaces of real numbers than to search through function spaces.

For the three specific configurations of neural networks that are considered in this thesis the theoretical foundations of the approximation capabilities are as follows. Hornik et al. (1989) show that 3-layer feed-forward networks with an activation function that is an arbitrary sigmoid function, a continuous, non-decreasing scalar function taking values in a bounded interval, can approximate any square integrable function (given sufficiently many hidden neurons); this is a generalization of the result of Gallant and White (1988),

who show this result for a specific sigmoid function (the 'cosine squasher'). The Type 1 network is a 3-layer network with sigmoid activation function.

Intuitively, one would expect the Type 2 network to have the same approximation capabilities as the Type 1 network (as long as one desires to approximate non-negative functions), as its output is the exponent of the output from a 3-layer network with sigmoidal activation. Moreover, Stinchcombe and White (1989,1990) show that activation functions in multi-layer networks do not have to be sigmoidal: they give sufficient conditions that are satisfied by a much larger class of functions than merely sigmoids (although for example not by polynomials).

The Type 3 network is a radial basis function (RBF) neural network. Many proofs of approximation properties for RBF networks can be found in literature. However, the restrictions $p_h \geq 0$, $(h = 1, \ldots, H)$, $\sum_{h=1}^{H} p_h = 1$ make that the Type 3 network does not automatically have all these RBF network capabilities. However, Zeevi and Meir (1997) show that under certain conditions any density function may be approximated to arbitrary accuracy by a convex combination of 'basis' densities; the mixture of Student t densities in (2.11) falls within their framework.

Table 2.2 gives an overview of the reasons for which we have chosen these particular specifications. The implications shown in this table will be clarified in the sequel of this chapter.

Throughout this thesis we use the term 'neural network' to denote the classes of functions described above; it should be mentioned here that in part of the literature, see *e.g.* Hastie et al. (2001), such methods are termed 'adaptive basis function methods' or 'dictionary methods'. A key ingredient of these methods is a search mechanism that constructs a linear combination of (nonlinear) basis functions that are chosen from a (possibly infinite) set or 'dictionary' of candidate basis functions.

Table 2.2: Motivation of the particular neural network specifications that are considered in this chapter

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

2.3.1 Constructing a neural network approximation to a density

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

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

$$SSR(A, b, c, d) = \sum_{i=1}^{N} \left(\tilde{p}(\theta^i) - nn\big(\theta^i \big| A, b, c, d\big)\right)^2, \tag{2.13}$$

where the notation c instead of C is used, since in our Type 1 and 2 networks this is a $(1 \times H_1)$ vector.[7] We choose the most parsimonious neural network, i.e. the one with the

[7]This optimization method, or learning process, is a *back-propagation* algorithm. In the neural network literature several algorithms are denoted by back-propagation. In these methods the derivatives of the *error function* (in our case the sum of squared residuals) with respect to the weights are analytically derived, where first the derivatives with respect to the last layer of weights is computed, after which one iteratively works backwards through the layers of the network using the chain-rule for partial derivatives in a clever way in order to minimize the required number of function evaluations.

least hidden neurons, 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 $\tilde{p}$ and nn at the points θ^i $(i = 1, \ldots, N)$.

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

In the case of a Type 1 (three-layer) neural network, we also have to deal with the problem that the neural network function is not automatically non-negative for each θ. In order to establish this a penalty term is added to (2.13), for example $-M \sum_{i=1}^{N} I\{nn(\theta^i) < 0\}\, nn(\theta^i)$ where M is a constant large enough to make nn positive (or only slightly negative) in all points θ^i $(i = 1, \ldots, N)$. Notice that if the minimum of $nn(\theta)$ is an (in absolute sense) very small negative value, one can simply subtract this negative value from the network's constant d, so that $nn(\theta)$ becomes non-negative for each θ. It should be mentioned that, since a neural network can have a surface that looks like a bed of nails, one should be very careful when checking the non-negativity. For example, one can look for the (global) minimum of $nn(\theta)$ by running a minimization procedure starting with several initial values. In our Type 2 (simplified four-layer) neural network the exponential function implies that non-negativity is automatically taken care of.

Type 3 (mixture of t) neural network approximation

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

First, compute the mode μ_1 and scale Σ_1 of the first Student t distribution in the mixture as $\mu_1 = \text{argmax}_\theta\ \tilde{p}(\theta)$, the mode of the target distribution, and Σ_1 as minus the inverse Hessian of $\log \tilde{p}(\theta)$ evaluated at its mode μ_1. Then draw a set of points θ^i $(i = 1, \ldots, N)$ from the 'first stage neural network' $nn(\theta) = t(\theta|\mu_1, \Sigma_1, \nu)$, with small ν to

allow for fat tails.[8] After that add components to the mixture, iteratively, by performing the following steps:

Step 1: Compute the importance sampling weights $w(\theta^i) = \tilde{p}(\theta^i)/nn(\theta^i)$ $(i = 1, \ldots, N)$. In order to determine the number of components H of the mixture we make use of a simple diagnostic criterion: the coefficient of variation, i.e. the standard deviation divided by the mean, of the IS weights $w(\theta^i)$ $(i = 1, \ldots, N)$. If the relative decrease in the coefficient of variation of the IS weights caused by adding one new Student-t component to the candidate mixture is small, e.g. less than 10%, then stop: the current $nn(\theta)$ is the Type 3 neural network approximation.[9] Otherwise, go to step 2.

Step 2: Add another Student t distribution with density $t(\theta|\mu_h, \Sigma_h, \nu)$ to the mixture with $\mu_h = \text{argmax}_\theta\ w(\theta) = \text{argmax}_\theta\{\tilde{p}(\theta)/nn(\theta)\}$ and Σ_h equal to minus the inverse Hessian of $\log w(\theta) = \log \tilde{p}(\theta) - \log nn(\theta)$ evaluated at its mode μ_h. Here $nn(\theta)$ denotes the mixture of $(h-1)$ Student t densities obtained in the previous iteration of the procedure. An obvious initial value for the maximization procedure for computing $\mu_h = \text{argmax}_\theta\ w(\theta)$ is the point θ^i with the highest weight $w(\theta^i)$ in the sample $\{\theta^i | i = 1, \ldots, N\}$. The idea behind this choice of μ_h and Σ_h is that the new t component should 'cover' a region where the weights $w(\theta)$ are relatively large: the point where the weight function $w(\theta)$ attains its maximum is an obvious choice for the mode μ_h, while the scale Σ_h is the covariance matrix of the local normal approximation to the distribution with density kernel $w(\theta)$ around the point μ_h.

If the region of integration of the parameters θ is bounded, it may occur that $w(\theta)$ attains its maximum at the boundary of the integration region; in this case minus the inverse Hessian of $\log w(\theta)$ evaluated at its mode μ_h may be a very poor scale matrix; in fact this matrix may not even be positive definite. In that case μ_h and Σ_h

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

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

are obtained as estimates of the mean and covariance matrix of a certain 'residual distribution' with density kernel:

$$res(\theta) = \max\{\tilde{p}(\theta) - \tilde{c}\, nn(\theta), 0\}, \tag{2.14}$$

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

$$w_{res}(\theta^i) = \frac{res(\theta^i)}{nn(\theta^i)} = \max\{w(\theta^i) - \tilde{c}, 0\} \quad \text{and} \quad \tilde{w}_{res}(\theta^i) = \frac{w_{res}(\theta^i)}{\sum_{i=1}^{N} w_{res}(\theta^i)}, \tag{2.15}$$

and μ_h and Σ_h are obtained as:

$$\mu_h = \sum_{i=1}^{N} \tilde{w}_{res}(\theta^i)\theta^i \qquad \Sigma_h = \sum_{i=1}^{N} \tilde{w}_{res}(\theta^i)(\theta^i - \mu_h)(\theta^i - \mu_h)'. \tag{2.16}$$

There are two issues relevant for the choice of $\tilde{c}$ in (2.14) and (2.15). First, the new t density should appear exactly at places where $nn(\theta)$ is too small (relative to $\tilde{p}(\theta)$), i.e. the scale should not be too large. Second, there should be enough points θ^i with $w(\theta^i) > \tilde{c}$ in order to make Σ_h nonsingular. A procedure is to calculate Σ_h for $\tilde{c}$ equal to 100 times the average value of $w(\theta^i)$ $(i = 1, \ldots, N)$; if Σ_h in (2.16) is nonsingular, accept $\tilde{c}$; otherwise lower $\tilde{c}$.

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^i 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:

$$E[w(\theta)^k] = \frac{1}{N} \sum_{i=1}^{N} \sum_{h=1}^{H} p_h\, w\big(\theta_h^i\big)^k \quad (k = 1, 2), \qquad w\big(\theta_h^i\big) = \frac{\tilde{p}(\theta_h^i)}{\sum_{l=1}^{H} p_l\, t(\theta_h^i|\mu_l, \Sigma_l, \nu)}. \tag{2.17}$$

Step 4: Draw a sample of N points θ^i $(i = 1, \ldots, N)$ from our new mixture of t distributions, $nn(\theta) = \sum_{h=1}^{H} p_h\, t(\theta|\mu_h, \Sigma_h, \nu)$, 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 N. The idea behind this is that the larger N is, the easier it is for the method to 'feel' the shape of the target density kernel, and to specify the t distributions of the mixture adequately.

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

2.3.2 Sampling from a neural network density

Type 1 (3-layer) neural network density

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

$$nn(\theta) = \sum_{h=1}^{H} \frac{c_h}{\pi} \arctan(a_h'\theta + b_h) + \frac{1}{2}\sum_{h=1}^{H} c_h + d, \tag{2.18}$$

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

Theorem 1: *The n-th integral $J_n(x)$ $(n = 1, 2, \ldots)$ of the arctangent function, $J_n(x) \equiv \int\cdots\int \arctan(x)dx\cdots dx$ with $x \in \mathbb{R}$, is given by*

$$J_n(x) = p_n(x)\arctan(x) + q_n(x)\ln(1+x^2) + r_n(x), \qquad x \in \mathbb{R}, \tag{2.19}$$

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

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

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

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

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

Proof: By induction; see appendix 2.A.1.[10]

A kernel of the cumulative distribution function of $\theta \sim nn(\theta)$ with nn in (2.18) is given by:

$$CDF_\theta(\theta_1, \ldots, \theta_k) = \left(\frac{1}{2} \sum_{h=1}^{H} c_h + d \right) (\theta_1 - \underline{\theta}_1) \cdots (\theta_k - \underline{\theta}_k) + \sum_{h=1}^{H} \frac{c_h}{\pi a_{h1} a_{h2} \cdots a_{hk}} \sum_{D_1=0}^{1} \cdots \sum_{D_k=0}^{1} (-1)^{D_1 + \cdots + D_k} J_k \left(\sum_{j=1}^{k} a_{hj} \theta_{j,D_j} + b_h \right), \quad (2.20)$$

where we define $\theta_{j,0} = \theta_j$ and $\theta_{j,1} = \underline{\theta}_j$ $(j = 1, 2, \ldots, k)$, the upper and lower bounds of the integration intervals; the primitive $J_k(\cdot)$ is given by (2.19) in Theorem 1.

The marginal distribution functions $CDF_{\theta_j}(\theta_j)$ $(j = 1, \ldots, k)$ are now obtained by taking $\theta_l = \bar{\theta}_l \; \forall l = 1, \ldots, k; l \neq j$ in (2.20). The conditional density kernel of $(\theta_1, \ldots, \theta_j)$ given $(\theta_{j+1}, \ldots, \theta_k)$ is simply obtained by substituting the values $\theta_{j+1}, \ldots, \theta_k$ into (2.18); a kernel of the conditional CDF is given by (2.20) with $\sum_{l=j+1}^{k} a_{hl}\theta_l + b_h$ instead of b_h (and j instead of k).

Sampling a random vector θ from the density kernel $nn(\theta)$ is easily done by drawing $U(0,1)$ variables and numerically inverting the distribution functions; it seems that taking a few steps of the bisection method followed by the Newton-Raphson method works well in practice. A more detailed description is given in appendix 2.A.2.

Type 2 (4-layer) neural network density

Suppose the joint density kernel of a certain $\theta \in \mathbb{R}^k$ is given by the Type 2 neural network:

$$nn(\theta) = \exp\left(\sum_{h=1}^{H} c_h \; plin(a_h'\theta + b_h) + d \right), \quad (2.21)$$

where each element θ_j is restricted to a certain finite interval $[\underline{\theta}_j, \bar{\theta}_j]$ $(j = 1, \ldots, k)$. It is easy to perform Gibbs sampling from this distribution, as one can divide the domain of each θ_j $(j = 1, \ldots, k)$ into a finite number of intervals on which the conditional neural network density is just the exponent of a linear function; the obvious reason for this is that a linear combination of piecewise-linear functions of θ_j is itself a piecewise-linear function of θ_j. Therefore we can analytically integrate the conditional neural network

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

density, and draw from it by analytically inverting the conditional CDF. Note that the three properties of g_2 mentioned below formula (2.10) are used here explicitly. A more detailed description of this procedure can be found in appendix 2.B.1.

Another possible method to draw from the Type 2 neural network density is auxiliary variable Gibbs sampling, which is a Gibbs sampling technique developed by Damien et al. (1999). The method is based on work of Edwards and Sokal (1988). In this method a vector of latent variables u is introduced in an artificial way in order to facilitate drawing from the full set of conditional distributions of θ_j $(j = 1, \ldots, k)$. In the case of our Type 2 neural network the vector of latent variables u is $(H \times 1)$ where conditionally on θ the u_h $(h = 1, \ldots, H)$ are independently drawn from uniform distributions:

$$u_h|\theta \;\sim\; U\left(0, \exp\left[c_h \, plin\left(\sum_{j=1}^{k} a_{hj}\theta_j + b_h\right)\right]\right), \qquad h = 1, \ldots, H. \tag{2.22}$$

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

$$\theta_{j,LB}(u, \theta_{-j}) \;=\; \max\left\{\underline{\theta}_j, \max_{1\le h\le H}\left\{\frac{1}{a_{hj}}\left(\frac{\log(u_h)}{c_h} - \frac{1}{2} - \left(\sum_{l=1, l\neq j}^{k} a_{hl}\theta_l + b_h\right)\right)\right| c_h a_{hj} > 0, 0 < \frac{\log(u_h)}{c_h} < 1\right\}\right\}, \tag{2.23}$$

$$\theta_{j,UB}(u, \theta_{-j}) \;=\; \min\left\{\bar{\theta}_j, \min_{1\le h\le H}\left\{\frac{1}{a_{hj}}\left(\frac{\log(u_h)}{c_h} - \frac{1}{2} - \left(\sum_{l=1, l\neq j}^{k} a_{hl}\theta_l + b_h\right)\right)\right| c_h a_{hj} < 0, 0 < \frac{\log(u_h)}{c_h} < 1\right\}\right\}. \tag{2.24}$$

The derivations of these conditional distributions are given in appendix 2.B.2. Using auxiliary variable Gibbs sampling, we do not have to restrict ourselves to the piecewise-linear function *plin* when specifying the activation function g_1; it allows for well-known activation functions such as the logistic and scaled arctangent functions.

Type 3 (mixture of t) neural network approximation

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

2.3.3 Neural network sampling algorithms

Once we have obtained a sample of random draws from the neural network density (kernel) $nn(\theta)$, we use this sample in order to estimate those characteristics of the target density (kernel) $\tilde{p}(\theta)$ that we are interested in. Two methods that we can use for this purpose are importance sampling and the (independence chain) Metropolis-Hastings algorithm, discussed in the introduction to this chapter. Note that in the case of a Type 2 (4-layer) neural network we need Gibbs sampling in order to obtain the sample, so that the consecutive draws are not independent. This case can be dealt with using a Metropolis-Hastings within Gibbs algorithm, in which a MH step is considered after each time an element θ_i is drawn from its conditional neural network distribution. So, we have the following eight 'neural network based' algorithms at hand:

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

Table 2.3 gives an overview.

Table 2.3: Overview of neural network based sampling algorithms

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

2.4 Comparison of performance of different neural networks

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)):

$$p(x_1, x_2) \propto \exp\left(-\frac{1}{2}\left[Ax_1^2x_2^2 + x_1^2 + x_2^2 - 2Bx_1x_2 - 2C_1x_1 - 2C_2x_2\right]\right), \tag{2.25}$$

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

$$X_1|X_2 = x_2 \sim N\left(\frac{3}{1+x_2^2}, \frac{1}{1+x_2^2}\right) \qquad X_2|X_1 = x_1 \sim N\left(\frac{3}{1+x_1^2}, \frac{1}{1+x_1^2}\right). \tag{2.26}$$

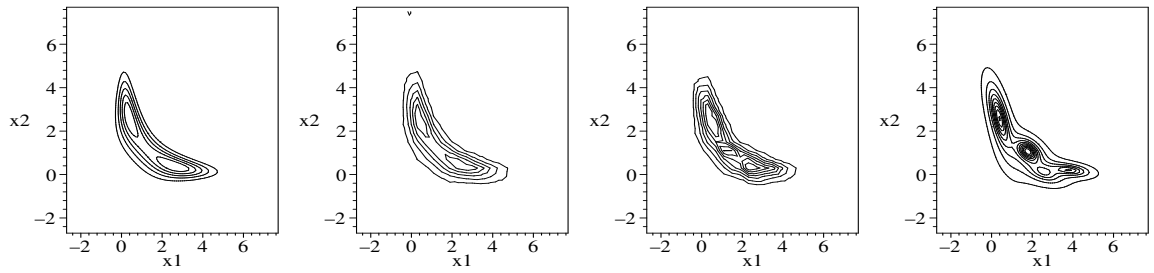

Figure 2.6: Contour plots: conditionally normal bivariate distribution in (2.26) (left), and its Type 1 (second), Type 2 (third), and Type 3 (right) neural network approximation

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

$$\{(X_1, X_2) | -2.5 \leq X_1 \leq 7.5, -2.5 \leq X_2 \leq 7.5\}. \tag{2.27}$$

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

The contourplots of the neural network approximations[11] are given by Figure 2.6, together with the contourplot of the target density. These contourplots confirm that the three classes of neural networks are able to provide reasonable approximations to the target density. Figure 2.7 illustrates how the AdMit procedure iteratively constructs an approximating (mixture of t) candidate density in four steps.

After we have constructed neural network approximations, we sample from these networks and use the samples in IS or the (independence chain) MH algorithm. Many diagnostic checks have been developed for assessing the convergence of the IS or MH method; see *e.g.* Geweke (1989) for the IS method and Cowles and Carlin (1996) and Brooks and Roberts (1998) for MCMC methods. Here we use the following simple heuristic rule to obtain estimates of the means with a precision of 1 decimal: for each algorithm we construct two samples, and we say that convergence has been achieved if the difference between the

[11] We constructed a Type 1 network with $H_1 = 50$ 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 $H_1 = 13$, $R^2 = 0.9944$ on its training set of 1000 points, and $R^2 = 0.9756$ on its test set of 5000 points; the $H_1 = 13$ 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 t distributions with a sample of 1000 IS weights with coefficient of variation equal to 0.840 (and in which the 5% most influential points have 11.6% weight).

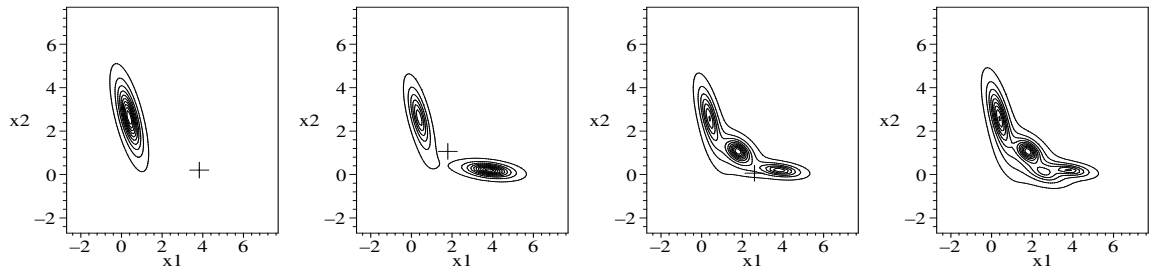

Figure 2.7: Illustration of the AdMit procedure for constructing a Type 3 (mixture of t) neural network approximation to a target (posterior) density: in four steps a candidate density is constructed; the cross denotes the point at which the weight function $p(x_1, x_2)/nn(x_1, x_2)$ corresponding to the displayed candidate density $nn(x_1, x_2)$ attains its maximum. For the four shown candidate densities the coefficient of variation of the importance sampling weights is 4.01, 1.39, 0.93, 0.87, respectively.

two estimates of $E(X_1)$ and the difference between the two estimates of $E(X_2)$ are both less than 0.05.[12] The results are in Table 2.4. 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 importance density coincide the RNE equals one, whereas a very poor importance density will have an RNE close to zero.

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

If we look at the computing times (on an AMD Athlon$^{\text{TM}}$ 1.4 GHz processor) required for generating the samples, we conclude that AdMit-IS and AdMit-MH are the winners in this example. In AdMit-IS or AdMit-MH the construction of the network, the sampling,

[12]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.

Table 2.4: Neural network based sampling results for the conditionally normal bivariate distribution in (2.26)

		Type 1 NN		Type 2 NN				Type 3 NN	
	real values	NNIS	NNMH	GiNNIS	GiNNMH	GiAuVa NNIS	GiAuVa NNMH	AdMit IS	AdMit MH
$E(X_1)$	1.459	1.487	1.504	1.472	1.433	1.468	1.477	1.464	1.467
(num. std. error)		(0.019)						(0.015)	
[RNE]		[0.896]						[0.649]	
$E(X_2)$	1.459	1.450	1.434	1.444	1.490	1.454	1.436	1.459	1.458
(num. std. error)		(0.019)						(0.016)	
[RNE]		[0.885]						[0.619]	
$\sigma(X_1)$	1.234	1.239	1.247	1.233	1.229	1.239	1.237	1.236	1.245
$\sigma(X_2)$	1.234	1.239	1.235	1.223	1.244	1.233	1.234	1.242	1.235
$\rho(X_1, X_2)$	-0.760	-0.764	-0.766	-0.755	-0.757	-0.758	-0.757	-0.759	-0.759
total time		257.0 s	257.0 s	66.5 s	79.9 s	81.3 s	85.6 s	2.0 s	2.0 s
time construction NN		225.2 s	225.2 s	62.6 s	62.6 s	62.6 s	62.6 s	1.1 s	1.1 s
time sampling		31.8 s	31.8 s	3.9 s	17.3 s	18.7 s	23.0 s	0.9 s	0.9 s
draws		5000	5000	10000	40000	80000	80000	10000	10000
time/draw		6.4 ms	6.4 ms	0.39 ms	0.43 ms	0.23 ms	0.29 ms	0.09 ms	0.09 ms
5% weights		6.3 %		7.2 %		7.2 %		12.9 %	
coeff. var. weights		0.382		0.239		0.251		0.840	
acc. rate			84.6%		90.0 %		92.7 %		52.7 %
serial corr. X_1			0.15	0.65	0.73	0.90	0.92		0.45
serial corr. X_2			0.14	0.67	0.72	0.84	0.86		0.45

and the IS or MH method require altogether 2.0 seconds, whereas the other methods take much more time to construct a network and to generate an adequate sample.

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

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

We conclude that the Type 3 (mixture of t) network clearly outperforms the other two networks, both in the construction and sampling process. The result that the optimization

(or learning) process of the Type 3 network, a Radial Basis Function (RBF) network, takes less time than those of the Type 1 and 2 networks, a Multi-Layer Perceptron (MLP) and a network whose output is the exponent of a MLP's output, is typical. The reason is that in RBF networks the optimization process is usually cleverly divided into different phases concerning different (groups of) weights, whereas in MLP's all weights are generally jointly optimized/learned.

The methods using Type 2 networks, especially the GiNNIS procedure (importance sampling using the Gibbs sampler without auxiliary variables), may be competitive if (much) better optimization techniques are used. Several different optimization methods than back-propagation 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. Application of optimization techniques that are specifically designed for neural network learning to the Type 1 and 2 networks is a topic for further research.

The possibility of including auxiliary variables in the Gibbs sampler of the Type 2 network may seem useless given the results of the experiment in this section. However, the auxiliary variable Gibbs sampler could also be used in a network whose output is the exponent of a certain 'full' 4-layer network (with multiple neurons in both the first and second hidden layer), using the piecewise-linear activation function in the first hidden layer and an arbitrary (analytically invertible) sigmoid (e.g. the logistic or arctangent function) in the second hidden layer. This possibility, which may be especially interesting when combined with specific neural network learning methods, is left as another topic for further research.

The Type 1 network has the interesting property that the integral of its functional form can be evaluated analytically. Next to that also the moments can be derived analytically, see appendix 2.A.3. 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. This approach is considered by Hoogerheide, Kaashoek and Van Dijk (2003b). However, in practice it will often be extremely difficult and/or time consuming to to find a network with almost perfect fit.

Still, an interesting question is whether a useful deterministic integration method could be based on this Type 1 neural network function; for example by dividing the region of integration into suitable subregions, constructing a Type 1 network that (approximately) interpolates the integrand on each subregion, and adding the analytically evaluated inte-

grals on the subregions. Note that the same kind of idea is the foundation of well-known deterministic integration rules such as the trapezoid rule and Simpson's rule, in which the integrals of interpolating polynomial for subregions are added. Notice that for the Type 1 network the integrand does not have to be a density; basically, it can be used to approximate any (square) integrable function.

2.5 Comparison of performance of neural networks with other methods

In the previous section the performance of the three types of neural networks was compared. The Type 3 neural network sampling method – in which an adaptive mixture of t densities (AdMit) was constructed – clearly outperformed the methods based on the other two methods. In this section we consider the posterior distributions in some instrumental variable (IV) regression models for simulated data in order to compare the performance of the AdMit procedure with some other sampling methods. Consider the following possibly overidentified IV model, also known as the incomplete simultaneous equations model (INSEM). Following Zellner et al. (1988), let:

$$y_1 = y_2\beta + \varepsilon \tag{2.28}$$

$$y_2 = X\pi + v \tag{2.29}$$

where y_1 is a $(T \times 1)$ vector of observations on the endogenous variable that is to be explained, y_2 is a $(T \times 1)$ vector of observations on the explanatory endogenous variable, X is a $(T \times k)$ matrix of weakly exogenous variables; β is a scalar structural parameter of interest, π is a $(k \times 1)$ vector of reduced form parameters. Assume that the rows of the matrix of error terms $(\varepsilon\ v)$ are independently normally distributed with (2×2) covariance matrix Σ with elements σ_{ij} $(i, j = 1, 2)$. The following non-informative prior density is specified:

$$p(\beta, \pi, \Sigma) \propto |\Sigma|^{-h/2} \text{ with } h > 0, \tag{2.30}$$

where the value $h = 3$ is chosen, which leads to the following joint posterior kernel of (β, π):

$$p(\beta, \pi | y_1, y_2, X) \propto \begin{vmatrix} (y_1 - y_2\beta)'(y_1 - y_2\beta) & (y_1 - y_2\beta)'(y_2 - X\pi) \\ (y_2 - X\pi)'(y_1 - y_2\beta) & (y_2 - X\pi)'(y_2 - X\pi) \end{vmatrix}^{-T/2}. \tag{2.31}$$

The IV model and the properties of the prior and posterior densities given above will be discussed more elaborately in chapters 4 and 5. Note that in the notation of the previous sections the parameters of interest are given by $\theta = (\beta, \pi)$; the number of elements of θ is $k + 1$ instead of k.

First, consider the joint posterior of π and β in (2.31) for a data set simulated from the model (2.28) - (2.29) with $k = 1$ instrument, $T = 100$ observations, 'true' values of parameters $\beta = 0$, $\pi = 0.1$ (weak identification), $\sigma_{11} = \sigma_{22} = 1$, $\sigma_{12} = 0.99$ (strong endogeneity); the elements of x are i.i.d. N(0,1) draws. This posterior density is truncated to the region[13]

$$\{(\pi, \beta)| - 0.25 \leq \pi \leq 0.25, -10 \leq \beta \leq 10\}. \tag{2.32}$$

The left panel of Figure 2.8 shows its contour plot on this region (2.32). The contour plot of the Type 3 neural network approximation[14] is given by the middle panel of this figure; this contour plot confirms that this class of neural networks is able to provide reasonable approximations to a wide class of (possibly multi-modal) target densities. In this example the Gibbs sampler failed: the Gibbs sequence remained in one of the two ridges for at least 100 million draws, yielding a scatter plot like the right panel of Figure 2.8. Of course, one can draw from the other ridge by choosing a different initial value, but it is not a trivial issue how to weight the results from the two ridges, i.e. it is not trivial to determine which part of the posterior probability mass is contained in each of both ridges.

Second, consider the joint posterior of $\pi = (\pi_1, \pi_2)'$ and β in (2.31) for $T = 50$ simulated data points from the model (2.28) - (2.29) with $\beta = 0$, $\sigma_{11} = \sigma_{22} = 1$, $\pi_1 = \pi_2 = 0.1$ (weak identification) and $\sigma_{12} = 0.99$ (strong endogeneity), with $k = 2$ vectors of instruments consisting of i.i.d. N(0,1) draws, truncated to the region

$$\{(\pi_1, \pi_2, \beta)| - 0.5 \leq \pi_i \leq 0.5 \ (i = 1, 2), -10 \leq \beta \leq 10\}. \tag{2.33}$$

Figure 2.9 shows the shape of a highest posterior density (HPD) credible set of (π_1, π_2, β) in the region (2.33) for this simulated data set.

We use our AdMit procedure to construct a Type 3 neural network approximation, a mixture of 15 Student t distributions (using $N = 5000$ points in the construction process), and use 1000000 draws from it in IS and MH; see Table 2.5. The reported

[13]For this exactly identified case, using the prior in (2.30), it is necessary to restrict the posterior density to a certain region, as otherwise the kernel in (2.31) corresponds to an improper density.

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

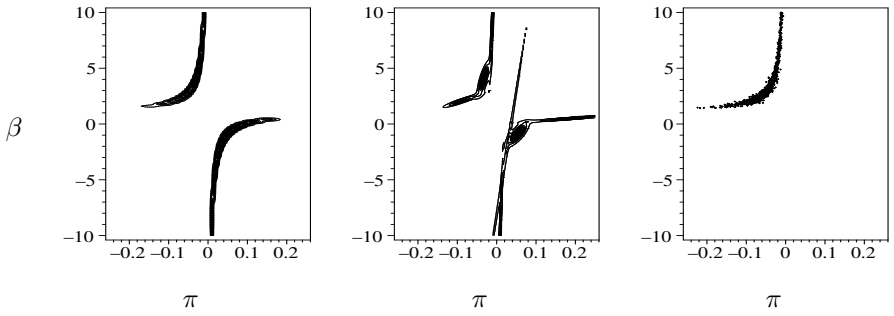

Figure 2.8: Contour plots in the $\pi \times \beta$ plane: joint posterior of π and β in IV model for simulated data set with $\pi = 0.1$, $\rho = 0.99$ (left), and its Type 3 neural network approximation (middle); scatter plot of sample obtained by the Gibbs sampler (right)

computing times correspond to an AMD Athlon^{TM} 1.4 GHz processor. We have repeated the algorithms 20 times; Table 2.5 shows the standard deviations of the 20 estimates of $E(\pi_1)$, $E(\pi_2)$ and $E(\beta)$. The table also shows numerical standard errors and the corresponding relative numerical efficiency (RNE), see Geweke (1989) (or the previous section for a short explanation).

The performance of AdMit-IS (in the same computing time) is compared with IS using a unimodal importance density, the Student t distribution with $\nu = 1$ degree of freedom. In order to give the unimodal density a fair chance, the mode and scale are first iteratively updated four times as the estimated mean and covariance matrix of the target distribution in the previous step. The results are in Table 2.5. AdMit-IS gives standard deviations of the 20 estimates of $E(\pi_1)$, $E(\pi_2)$, $E(\beta)$ that are 2.3, 2.0, 2.2 times as small, respectively, while the numerical standard errors are 1.9, 1.9, 3.4 times as small for AdMit-IS. Also notice the huge differences between the RNEs (especially for the estimate of $E(\beta)$), the total weights of the 5% most influential points and the coefficients of variation of the weights in the two IS methods.

We compare the performance of AdMit-MH with the independence chain MH algorithm using a Student t distribution with $\nu = 1$ degree of freedom, and with the random walk (RW) Metropolis-Hastings algorithm with candidate steps from a t_1 distribution. The scale (and mode) are first iteratively updated 4 times as the estimated covariance matrix (and mean) of the target distribution in the previous step. The results are in Table 2.5. AdMit-MH yields standard deviations of the 20 estimates of $E(\pi_1)$, $E(\pi_2)$, $E(\beta)$ that are 1.9, 1.9, 3.6 times smaller than t_1 (independence chain) MH, and 1.6, 1.5, 1.3 times

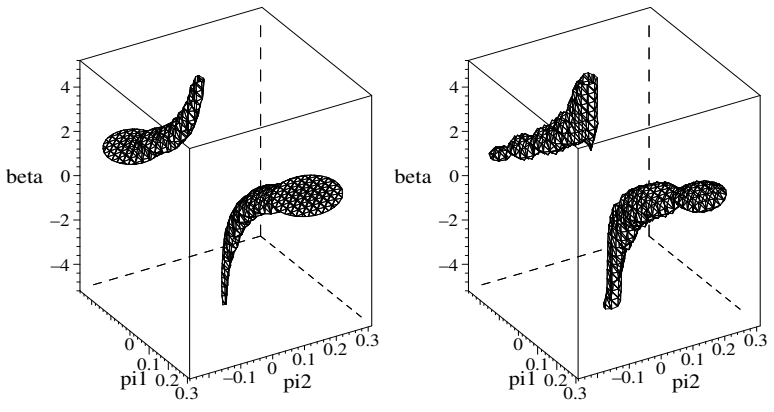

Figure 2.9: HPD credible set for parameters π_1, π_2, β in IV model (2.28) - (2.29) for a simulated data set from this model with strong endogeneity ($\rho = 0.99$) combined with weak identification ($\pi_1 = \pi_2 = 0.1$; left), and a 'highest candidate density' set for the mixture-of-t candidate density (right).

smaller than RW MH. Also note that AdMit-MH has much higher acceptance rate and lower (first order) serial correlations in the MH chain.

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

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

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

Finally, consider the joint posterior of π and β in (2.31) for a data set simulated from the model (2.28) - (2.29), with $k = 1$ instrument, similar to the first simulated data set in this section; however, with 'true' values of parameters $\pi = 1$ (strong identification) and $\sigma_{12} = 0$ (no endogeneity). The posterior in (2.31) is truncated to the region

$$\{(\pi, \beta)| - 0.5 \leq \pi \leq 1.5, -10 \leq \beta \leq 10\}. \tag{2.34}$$

Figure 4.1 shows its contour plot, which shows an elliptical shape. We construct a Type 3 neural network approximation, a mixture of two Student t distributions. The same simple heuristic rule as in the previous section is used to obtain estimates of the means

Table 2.5: Sampling results for the non-elliptically shaped posterior distribution in the IV regression (2.28) - (2.29) with $k = 2$ instruments for simulated data with $\pi = (0.1, 0.1)'$ (weak identification), $\sigma_{12} = 0.99$ (strong endogeneity)

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

Table 2.6: Sampling results for the elliptically shaped posterior distribution in the IV regression (2.28) - (2.29) with $k = 1$ instrument for simulated data with $\pi = 1$ (strong identification) and $\sigma_{12} = 0$ (no endogeneity)

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

with (roughly) a precision of 2 decimals: for each algorithm we construct two samples, and we say that convergence has been achieved if the difference between the two estimates of $E(\pi)$ and the difference between the two estimates of $E(\beta)$ are both less than 0.005.[15] The results are in Table 2.6. We compare AdMit's performance with the Gibbs sampler, the random walk MH algorithm with candidate steps from a t_1 distribution with scale matrix equal to minus the inverse Hessian of the log-posterior kernel evaluated at its mode, and IS/MH with a t_1 or normal candidate density around the mode of the target distribution. In this case of an elliptical target distribution the Gibbs sampler and the methods using a unimodal candidate density all perform well. Although the neural network approach is feasible in this example, it is slower than several competing algorithms. This emphasizes that different sampling methods dominate in different cases; the neural network approach is especially useful for target densities with highly non-elliptical contours.

2.6 Conclusions

Evaluating integrals is a crucial step in Bayesian inference. In case one desires or needs to use importance sampling (IS) or the (independence chain) Metropolis-Hastings (MH) algorithm, it is important that the target (posterior) density is roughly mimicked by the candidate density. If the target distribution is highly non-elliptical (e.g. displaying multimodality), a unimodal, elliptical candidate distribution such as the normal or Student-t distribution may yield results very slowly or may even be unreliable in the sense that certain modes are completely 'missed'.

In order to perform IS or the MH algorithm in cases of highly non-elliptical target (posterior) distributions, we have introduced a class of neural network sampling algorithms. In these algorithms neural network functions are used as an importance or candidate density in IS or the MH algorithm. Neural networks are natural importance or candidate densities, as they have a universal approximation property and are easy to sample from. We have shown how to sample from three types of neural networks. One can sample directly from a certain 3-layer network. Using a 4-layer network one can, depending on the specification of the network, either use a Gibbs sampling approach or sample directly from a mixture of distributions. A key step in the proposed class of methods is the

[15] Like in the example of the previous section, 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.

construction of a neural network that approximates the target density accurately. The methods have been tested on an illustrative example; the 4-layer network specified as the mixture of t distributions performed the best among the proposed sampling procedures. In another experiment concerning a bimodal posterior distribution in an IV regression for a simulated data set the approach using a mixture of t distributions provided (in the same computing time) more accurate results than IS with a unimodal importance density or a random walk Metropolis-Hastings algorithm, whereas the Gibbs sampler failed in this example. These results indicate the feasibility and the possible usefulness of the neural network approach.

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

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

Third, more experience is needed with empirical econometric models like the models of local average treatment effects, see Imbens and Angrist (1994), or the business cycle models as specified by Hamilton (1989) and Paap and Van Dijk (2003), or stochastic volatility

models as given by Shephard (1996), and dynamic panel data models; see Pesaran and Smith (1995).

Fourth, the neural network approximations proposed in this chapter may be useful for modelling such processes as volatility in financial series, see *e.g.* Donaldson and Kamstra (1997), and for evaluating option prices, see Hutchinson, Lo and Poggio (1994).

2.A Derivations for Type 1 (3-layer) neural network

Appendix 2.A.1 gives analytical expressions for the integrals of the arctangent function. Appendix 2.A.2 shows how these expressions are used in order to sample from the Type 1 (3-layer) neural network distribution. In appendix 2.A.3 these expressions are used to obtain analytical expressions for the moments of the Type 1 (3-layer) neural network distribution.

2.A.1 Analytical expression for the integrals of the arctangent function

Theorem A.1: The n-th integral of the arctangent function $J_n(x)$

$$J_n(x) \equiv \int \cdots \int \arctan(x) dx \cdots dx$$

is given by

$$J_n(x) = p_n(x) \arctan(x) + q_n(x) \ln(1 + x^2) + r_n(x), \tag{2.35}$$

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

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

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

$$p_{n,k} = \begin{cases} \frac{(-1)^{(n-k)/2}}{(n-k)!k!} & \text{if } n-k \text{ is even} \\ \\ 0 & \text{if } n-k \text{ is odd} \end{cases} \qquad q_{n,k} = \begin{cases} \frac{(-1)^{(n-k+1)/2}}{2(n-k)!k!} & \text{if } n-k \text{ is odd} \\ \\ 0 & \text{if } n-k \text{ is even} \end{cases} \tag{2.36}$$

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

Proof: We will prove this theorem by induction. First, note that for $n = 1$ the proposition holds, as we have by partial integration:

$$\int \arctan(x)dx = x\ \arctan(x) - \frac{1}{2}\ln(1+x^2), \tag{2.37}$$

Now suppose that our proposition holds for a certain positive integer n. Then we have to show that this implies that the proposition also holds for $n+1$.

First, note that for any non-negative integer k partial integration yields:

$$\begin{aligned} \int x^k \arctan(x)dx &= \frac{1}{k+1}x^{k+1}\arctan(x) - \frac{1}{k+1}\int \frac{x^{k+1}}{1+x^2}dx, \\ \int x^k \ln(1+x^2)dx &= \frac{1}{k+1}x^{k+1}\ln(1+x^2) - \frac{2}{k+1}\int \frac{x^{k+2}}{1+x^2}dx. \end{aligned} \tag{2.38}$$

Second, notice that a partial fraction decomposition yields:

$$\int \frac{x^m}{1+x^2}dx = \begin{cases} (-1)^{m/2}\ \arctan(x) + \sum_{i=0}^{(m-2)/2} \frac{(-1)^i}{m-1-2i}\, x^{m-1-2i} & \text{if } m \text{ is even,} \\ (-1)^{(m-1)/2}\, \frac{\ln(1+x^2)}{2} + \sum_{i=0}^{(m-3)/2} \frac{(-1)^i}{m-1-2i}\, x^{m-1-2i} & \text{if } m \text{ is odd.} \end{cases} \tag{2.39}$$

We may omit the polynomials in (2.39), since these would eventually be absorbed by the irrelevant polynomial r_n in formula (2.35), anyway. The induction assumption is that for a certain n it holds that:

$$\begin{aligned} J_n(x) &= (p_{n,0} + p_{n,1}\,x + \ldots + p_{n,n}\,x^n)\arctan(x) \\ &\quad + \left(q_{n,0} + q_{n,1}\,x + \ldots + q_{n,n-1}\,x^{n-1}\right)\ln(1+x^2) \end{aligned} \tag{2.40}$$

where the coefficients $p_{n,k}$ $(k = 0, 1, \ldots, n)$ and $q_{n,k}$ $(k = 0, 1, \ldots, n-1)$ are given by (2.36). It follows from (2.38) and (2.39) that:

$$\begin{aligned} J_{n+1}(x) &= \int J_n(x)dx \\ &= \left(p_{n+1,0} + p_{n,0}\,x + \frac{p_{n,1}}{2}\,x^2 + \ldots + \frac{p_{n,n}}{n+1}\,x^{n+1}\right)\arctan(x) \\ &\quad + \left(q_{n+1,0} + q_{n,0}\,x + \frac{q_{n,1}}{2}\,x^2 + \ldots + \frac{q_{n,n-1}}{n}\,x^n\right)\ln(1+x^2) \end{aligned}$$

Note that $J_{n+1}(x)$ has the shape of formula (2.35) with $p_{n+1,k} = p_{n,k-1}/k$ $(k = 1, \ldots, n+1)$ and $q_{n+1,k} = q_{n,k-1}/k$ $(k = 1, \ldots, n)$. Combining this with the induction assumption, it is easy to see the validity of the formulas for $p_{n+1,k}$ and $q_{n+1,k}$ for $k \geq 1$. Now we only have to prove that $p_{n+1,0}$ and $q_{n+1,0}$ are also given by (2.36). From (2.38) and (2.39) we have:

$$p_{n+1,0} = \sum_{\{k|1\leq k\leq n;k \text{ odd}\}} -\frac{(-1)^{(k+1)/2}}{k+1}\ p_{n,k} + \sum_{\{k|0\leq k\leq n-1;k \text{ even}\}} -\frac{2(-1)^{(k+2)/2}}{k+1}\ q_{n,k}. \tag{2.41}$$

If n is even, all $p_{n,k}$'s and $q_{n,k}$'s in the two summations of (2.41) are equal to zero, so that in that case $p_{n+1,0} = 0$. If n is odd, we have:

$$p_{n+1,0} = \sum_{\{k|1\leq k\leq n;k \text{ odd}\}} -\frac{(-1)^{(n+1)/2}}{(n-k)!(k+1)!} + \sum_{\{k|0\leq k\leq n-1;k \text{ even}\}} -\frac{(-1)^{(n+3)/2}}{(n-k)!(k+1)!}, \quad (2.42)$$

which can be rewritten as:

$$p_{n+1,0} = \frac{(-1)^{(n+1)/2}}{(n+1)!}\sum_{k=0}^{n}(-1)^k \left(\begin{array}{c} n+1 \\ k+1 \end{array} \right) = \frac{(-1)^{(n+1)/2}}{(n+1)!}, \quad (2.43)$$

where the last equality of (2.43) follows from Newton's binomium. The proof for $q_{n+1,0}$ is similar. We conclude that $p_{n+1,0}$ and $q_{n+1,0}$ are also given by (2.36), so that we have proved the theorem by induction. Q.E.D.

2.A.2 Marginal and conditional CDF of the Type 1 (3-layer) neural network density

Suppose the random vector $X = (X_1, \ldots, X_n)'$ has the following density $p(x_1, \ldots, x_n)$:

$$p(x_1, \ldots, x_n) = \begin{cases} nn(x_1, \ldots, x_n) & \text{if } \underline{x}_i \leq x_i \leq \bar{x}_i \quad \forall\, i = 1, \ldots, n \\ \\ 0 & \text{else} \end{cases} \quad (2.44)$$

where $[\underline{x}_i, \bar{x}_i]$ is the interval to which the variable x_i $(i = 1, 2, \ldots, n)$ is restricted, and where $nn(x_1, \ldots, x_n)$ is the following three-layer neural network function:

$$nn(x_1, \ldots, x_n) = \sum_{h=1}^{H} \frac{c_h}{\pi} \arctan(a_h' x + b_h) + \frac{1}{2}\sum_{h=1}^{H} c_h + d. \quad (2.45)$$

Then the cumulative distribution function of X is given by:

$$\begin{aligned} CDF_X(\tilde{x}_1, \ldots, \tilde{x}_n) &= \int_{\underline{x}_n}^{\tilde{x}_n} \cdots \int_{\underline{x}_2}^{\tilde{x}_2} \int_{\underline{x}_1}^{\tilde{x}_1} nn(x_1, \ldots, x_n) dx_1 dx_2 \cdots dx_n \\ &= \sum_{h=1}^{H} \frac{c_h}{\pi} \int_{\underline{x}_n}^{\tilde{x}_n} \cdots \int_{\underline{x}_2}^{\tilde{x}_2} \int_{\underline{x}_1}^{\tilde{x}_1} \arctan(a_h' x + b_h) dx_1 dx_2 \cdots dx_n \\ &\quad + \left(\frac{1}{2}\sum_{h=1}^{H} c_h + d \right) x_1 x_2 \cdots x_n. \end{aligned} \quad (2.46)$$

Using the fact that $dx_1 = d(a_h'x + b_h)/a_{h1}$ (for constant values of $x_2, \ldots, x_n$), we make the following change of variables:

$$\begin{aligned}\int_{\underline{x}_1}^{\tilde{x}_1} \arctan(a_h'x + b_h)dx_1 &= \frac{1}{a_{h1}} \int_{a_{h1}\underline{x}_1 + a_{h,-1}'x_{-1} + b_h}^{a_{h1}\tilde{x}_1 + a_{h,-1}'x_{-1} + b_h} \arctan(a_h'x + b_h)d(a_h'x + b_h) \\ &= \frac{1}{a_{h1}} \left[J_1(a_{h1}\tilde{x}_1 + a_{h,-1}'x_{-1} + b_h) - J_1(a_{h1}\underline{x}_1 + a_{h,-1}'x_{-1} + b_h) \right],\end{aligned}$$

where we define $a_{h,-1} = (a_{h2}, \ldots, a_{hn})'$ and $x_{-1} = (x_2, \ldots, x_n)'$. If we continue in this way, we obtain the following formula:

$$\begin{aligned}&\int_{\underline{x}_n}^{\tilde{x}_n} \cdots \int_{\underline{x}_2}^{\tilde{x}_2} \int_{\underline{x}_1}^{\tilde{x}_1} \arctan(a_h'x + b_h)dx_1 dx_2 \cdots dx_n = \qquad (2.47) \\ &= \frac{1}{a_{h1}a_{h2}\cdots a_{hn}} \sum_{D_1=0}^{1} \cdots \sum_{D_n=0}^{1} (-1)^{D_1+D_2+\cdots+D_n} J_n(a_{h1}x_{1,D_1} + \cdots + a_{hn}x_{n,D_n} + b_h)\end{aligned}$$

where we define $x_{i,0} = \tilde{x}_i$ and $x_{i,1} = \underline{x}_i$ $(i = 1, 2, \ldots, n)$, the upper and lower bounds of the integration intervals. The primitive $J_n(x)$ is given by Theorem A.1 in appendix A.1. Substituting (2.47) into (2.46) yields:

$$CDF_x(\tilde{x}_1, \ldots, \tilde{x}_n) = \left(\frac{1}{2} \sum_{h=1}^{H} c_h + d \right) x_1 x_2 \cdots x_n +$$

$$+ \sum_{h=1}^{H} \frac{c_h}{\pi a_{h1}a_{h2}\cdots a_{hn}} \sum_{D_1=0}^{1} \cdots \sum_{D_n=0}^{1} (-1)^{D_1+\cdots+D_n} J_n \left(\sum_{i=1}^{n} a_{hi}x_{i,D_i} + b_h \right). \qquad (2.48)$$

The marginal distribution functions $CDF_{X_j}(x_j)$ $(j = 1, \ldots, n)$ are now obtained by taking $\tilde{x}_i = \bar{x}_i \ \forall i = 1, \ldots, n; i \neq j$:

$$CDF_{X_j}(x_j) = CDF_x(\bar{x}_1, \ldots, \bar{x}_{j-1}, x_j, \bar{x}_{j+1}, \ldots, \bar{x}_n). \qquad (2.49)$$

The conditional CDF of X_j given $X_{j+1}, \ldots, X_n$ is derived in a similar way, simply by substituting $\sum_{i=j+1}^{n} a_{hi}x_i + b_h$ for b_h and treating the neural network as a function of $x_1, \ldots, x_j$.

As we have explicit formulas for the marginal and conditional distribution functions, it is easy to sample a random vector from a three-layer neural network density with (scaled) arctangent activation function. We can use the numerical inverse transformation method in the following way:

Step 1: Draw n independent U(0,1) variables $U_1, U_2, \ldots, U_n$.

Step 2: Draw X_n from its marginal distribution by computing the value of X_n such that $CDF_{X_n}(X_n) = U_n$ (using, for example, the bisection method).

Step 3: For $j = n-1, n-2, \ldots, 1$ iteratively draw X_j from its conditional distribution on $X_{j+1}, \ldots, X_n$ by computing the value of X_j such that $CDF(X_j | X_{j+1}, \ldots, X_n) = U_j$.

2.A.3 Moments of the Type 1 (3-layer) neural network distribution

Suppose the vector $X = (X_1, \ldots, X_n)'$ has the three-layer neural network density $p(x_1, \ldots, x_n)$ given by (2.44) and (2.45). Then the expectation of X_n^k $(k = 1, 2, \ldots)$ is given by:

$$\begin{aligned} & E(X_n^k) = \\ = & \int_{\underline{x}_n}^{\bar{x}_n} \int_{\underline{x}_{n-1}}^{\bar{x}_{n-1}} \cdots \int_{\underline{x}_1}^{\bar{x}_1} x_n^k \; nn(x_1, \ldots, x_n) dx_1 \cdots dx_{n-1} dx_n \\ = & \sum_{h=1}^{H} \frac{c_h}{\pi a_{h1} \cdots a_{h,n-1}} \sum_{D_1=0}^{1} \cdots \sum_{D_{n-1}=0}^{1} \Big[(-1)^{D_1 + \cdots + D_{n-1}} \times \\ & \qquad \times \int_{\underline{x}_n}^{\bar{x}_n} x_n^k \, J_{n-1} \left(\sum_{i=1}^{n-1} a_{hi} x_{i,D_i} + a_{hn} x_n + b_h \right) dx_n \Big] \\ & + \left(\frac{1}{2} \sum_{h=1}^{H} c_h + d \right) \frac{1}{k+1} (\bar{x}_1 - \underline{x}_1) \cdots (\bar{x}_{n-1} - \underline{x}_{n-1}) (\bar{x}_n^{k+1} - \underline{x}_n^{k+1}), \end{aligned} \tag{2.50}$$

where we define $x_{i,0} = \bar{x}_i$ and $x_{i,1} = \underline{x}_i$ $(i = 1, 2, \ldots, n-1)$, the upper and lower bounds of the integration intervals. We now make use of the following theorem:

Theorem A.2: If the n-th integral of a certain function $f : \mathbb{R} \to \mathbb{R}$ is given by $J_n : \mathbb{R} \to \mathbb{R}$, then it holds for $a_h, x \in \mathbb{R}^n$, $b_h \in \mathbb{R}$ and $k = 0, 1, 2, \ldots$ that:

$$\int x_i^k \, J_n(a_h' x + b_h) dx_i = \frac{1}{a_{hi}} \sum_{m=0}^{k} \left(-\frac{1}{a_{hi}} \right)^m \frac{k!}{(k-m)!} \, x_i^{k-m} \, J_{n+1+m}(a_h' x + b_h). \tag{2.51}$$

Proof: We will prove this theorem by induction with respect to k. First, note that for $k = 0$ we have:

$$\int J_n(a_h' x + b_h) dx_i = \frac{1}{a_{hi}} \int J_n(a_h' x + b_h) d(a_h' x + b_h) = \frac{1}{a_{hi}} J_{n+1}(a_h' x + b_h),$$

which clearly corresponds to Theorem A.2 for $k = 0$. Now suppose that our proposition holds for a certain nonnegative integer k. Then we have to show that this implies that the proposition also holds for $k + 1$.

Partial integration with x_i^{k+1} as the factor to be differentiated yields:

$$\int x_i^{k+1}\, J_n(a_h'x+b_h)dx_i = x_i^{k+1}\,\frac{1}{a_{hi}}\,J_{n+1}(a_h'x+b_h) - \frac{k+1}{a_{hi}}\int x_i^k\, J_{n+1}(a_h'x+b_h)dx_i. \quad (2.52)$$

The induction assumption is that Theorem A.2 holds for the value k. Using this induction assumption we rewrite the second term of (2.52) as:

$$\begin{aligned} & -\frac{1}{a_{hi}}\,(k+1)\int x_i^k\, J_{n+1}(a_h'x+b_h)dx_i = \\ = \;& \frac{1}{a_{hi}}\sum_{j=1}^{k+1}\left(-\frac{1}{a_{hi}}\right)^j \frac{(k+1)!}{(k+1-j)!}\, x_i^{k+1-j}\, J_{n+1+j}(a_h'x+b_h) \end{aligned} \quad (2.53)$$

Adding (2.53) to the first term of (2.52) yields:

$$\int x_i^{k+1}\, J_n(a_h'x+b_h)dx_i = \frac{1}{a_{hi}}\sum_{j=0}^{k+1}\left(-\frac{1}{a_{hi}}\right)^j \frac{(k+1)!}{(k+1-j)!}\, x_i^{k+1-j}\, J_{n+1+j}(a_h'x+b_h)$$

which is just equation (2.51) with $k+1$ instead of k. We conclude that we have proved Theorem A.2 by induction. Q.E.D.

Substituting equation (2.51) of Theorem A.2 into (2.50) now yields $E(X_n^k)$, which can be easily adjusted to the general case of $E(X_i^k)$ $(i = 1, 2, \ldots, n)$ by taking a_{hi} and x_i instead of a_{hn} and x_n:

$$\begin{aligned} E(X_i^k) \;=\; & \sum_{h=1}^{H}\frac{c_h}{\pi a_{h1}\cdots a_{hn}}\sum_{D_1=0}^{1}\cdots\sum_{D_n=0}^{1}\Big[(-1)^{D_1+\cdots+D_n}\times \\ & \times\sum_{m=0}^{k}\left(-\frac{1}{a_{hi}}\right)^m\frac{k!}{(k-m)!}\, x_i^{k-m}\, J_{n+m}\left(\sum_{i=1}^{n}a_{hi}x_{i,D_i}+b_h\right)\Big] \\ & +\left(\frac{1}{2}\sum_{h=1}^{H}c_h+d\right)\frac{1}{k+1}(\bar{x}_i^{k+1}-\underline{x}_i^{k+1})\prod_{j=1;j\neq i}^{n}(\bar{x}_j-\underline{x}_j) \end{aligned} \quad (2.54)$$

In a similar fashion it can be derived that $E(X_iX_j)$ $(i,j=1,2,\ldots,n; i\neq j)$ is equal to:

$$\begin{aligned} E(X_iX_j) &= \sum_{h=1}^{H}\frac{c_h}{\pi a_{h1}\cdots a_{hn}}\sum_{D_1=0}^{1}\cdots\sum_{D_n=0}^{1}(-1)^{D_1+\cdots+D_n}\times \\ &\quad\times\left[x_ix_jJ_n\left(\sum_{i=1}^{n}a_{hi}x_{i,D_i}+b_h\right)\right. \\ &\quad -\frac{a_{hi}x_i+a_{hj}x_j}{a_{hi}a_{hj}}J_{n+1}\left(\sum_{i=1}^{n}a_{hi}x_{i,D_i}+b_h\right) \\ &\quad\left.+\frac{1}{a_{hi}a_{hj}}J_{n+2}\left(\sum_{i=1}^{n}a_{hi}x_{i,D_i}+b_h\right)\right] \\ &\quad+\left(\frac{1}{2}\sum_{h=1}^{H}c_h+d\right)\frac{1}{4}(\bar{x}_i^2-\underline{x}_i^2)(\bar{x}_j^2-\underline{x}_j^2)\prod_{k=1;k\neq i,j}^{n}(\bar{x}_k-\underline{x}_k). \end{aligned} \tag{2.55}$$

Using formulas (2.54) and (2.55), one can easily compute statistics of a three-layer feed-forward neural network distribution, such as mean, variance, skewness, kurtosis, covariances and correlations.

2.B Derivations for Type 2 (4-layer) neural network

Appendix 2.B.1 discusses how to draw from the Type 2 (4-layer) neural network distribution using Gibbs sampling. Appendix 2.B.2 shows another way to draw from the Type 2 (4-layer) neural network: auxiliary variable Gibbs sampling.

2.B.1 Gibbs sampling from the Type 2 (4-layer) neural network distribution

Suppose a density kernel of $X\in\mathbb{R}^n$ is given by

$$p(x)=\begin{cases} nn(x) & \text{if } x_i\in[\underline{x}_i,\bar{x}_i]\ \ \forall i=1,\ldots,n \\ 0 & \text{else} \end{cases} \tag{2.56}$$

where $[\underline{x}_i,\bar{x}_i]$ is the interval to which X_i $(i=1,\ldots,n)$ is restricted. Suppose the function $nn(x)$ corresponds to the following four-layer feed-forward neural network with n inputs x_i $(i=1,\ldots,n)$, and H hidden neurons:

$$nn(x)=\exp\left(\sum_{h=1}^{H}c_h\,plin\left(\sum_{i=1}^{n}a_{hi}x_i+b_h\right)+d\right), \tag{2.57}$$

where $plin : \mathbb{R} \to \mathbb{R}$ is the following piecewise-linear function:

$$plin(x) = \begin{cases} 0 & x < -1/2 \\ x + 1/2 & -1/2 \leq x \leq 1/2 \\ 1 & x > 1/2 \end{cases} \tag{2.58}$$

We rewrite the neural network density $nn(x) = nn(x_j, x_{-j})$ as

$$nn(x_j, x_{-j}) \quad \propto \quad \exp\left(\sum_{h=1}^{H} c_h \, plin\left(a_{hj} x_j + \sum_{i=1, i\neq j}^{n} a_{hi} x_i + b_h\right)\right),$$

which is a kernel of the conditional density of x_j given x_{-j}. For each hidden neuron h ($h = 1, \ldots, H$) there are two points x_j where its input $a_h' x + b_h$ moves from one of the intervals $(-\infty, -1/2)$, $[-1/2, 1/2]$ and $(1/2, \infty)$ to another one:

$$a_{hj} x_j + \sum_{i=1, i\neq j}^{n} a_{hi} x_i + b_h = \pm\frac{1}{2} \quad \Leftrightarrow \quad x_j = \frac{1}{a_{hj}}\left(\pm\frac{1}{2} - \sum_{i=1, i\neq j}^{n} a_{hi} x_i - b_h\right). \tag{2.59}$$

Consider only those 'changing points' $\tilde{x}_{j,k}$ ($k = 1, \ldots, m$ with $m \leq 2H$) that are in the interval of interest $[\underline{x}_j, \bar{x}_j]$, and order these m points such that:

$$\tilde{x}_{j,1} < \tilde{x}_{j,2} < \cdots < \tilde{x}_{j,m-1} < \tilde{x}_{j,m}$$

If we define $\tilde{x}_{j,0} = \underline{x}_j$ and $\tilde{x}_{j,m+1} = \bar{x}_j$, we have $m+1$ intervals $[\tilde{x}_{j,k}, \tilde{x}_{j,k+1}]$ ($k = 0, 1, \ldots, m$) on which a kernel of the conditional density of X_j given X_{-j} is given by:

$$nn(x_j, x_{-j}) \propto \exp(\tilde{a}_k x_j + \tilde{b}_k) \tag{2.60}$$

for certain constants $\tilde{a}_k$ and $\tilde{b}_k$ ($k = 0, 1, \ldots, m$). The primitive of (2.60) is given by

$$\int \exp(\tilde{a}_k x_j + \tilde{b}_k) dx_j = \begin{cases} \frac{1}{\tilde{a}_k} \exp(\tilde{a}_k x_j + \tilde{b}_k) + C_k & \text{if } \tilde{a}_k \neq 0 \\ \\ \exp(\tilde{b}_k) x_j + C_k & \text{if } \tilde{a}_k = 0. \end{cases}$$

where C_k ($k = 0, 1, \ldots, m$) are integration constants that we specify in such a way that the CDF starts at the value 0 and is continuous in x_j. After this kernel of the conditional CDF has been obtained, X_j is drawn from its conditional distribution using the inverse transformation method: one draws $U \sim U(0, 1)$ and computes:

$$X_j = \frac{\log\left[\tilde{a}_k \left(S\,U - C_k\right)\right] - \tilde{b}_k}{\tilde{a}_k} \qquad \text{or} \qquad X_j = \frac{S\,U - C_k}{\exp(\tilde{b}_k)}$$

depending on whether X_j falls in a region with $\tilde{a}_k = 0$ or not; S is the 'scaling constant' of the kernel, which is computed as the value of the kernel of the conditional CDF at $\bar{x}_j$.

Since it is easy to draw X_j conditional on X_{-j} ($j = 1, \ldots, n$), it is easy to perform Gibbs sampling from the Type 2 (4-layer) neural network distribution.

2.B.2 Auxiliary variable Gibbs sampling from the Type 2 (4-layer) neural network distribution

Suppose a density kernel of $X \in \mathbb{R}^n$ is given by

$$p(x) = \begin{cases} nn(x) & \text{if } x_i \in [\underline{x}_i, \bar{x}_i] \ \ \forall i = 1, \ldots, n \\ 0 & \text{else} \end{cases} \tag{2.61}$$

where $[\underline{x}_i, \bar{x}_i]$ is the interval to which X_i $(i = 1, \ldots, n)$ is restricted. Suppose the function $nn(x)$ corresponds to the following four-layer feed-forward neural network with n inputs x_i $(i = 1, \ldots, n)$, and H hidden neurons:

$$nn(x) = \exp\left(\sum_{h=1}^{H} c_h \, g\left(\sum_{i=1}^{n} a_{hi} x_i + b_h\right) + d\right), \tag{2.62}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a monotonically increasing function taking its values in [0,1], which is invertible on the interval $(\underline{x}, \bar{x})$ where it takes its values in (0,1). We will denote this invertible function by $\tilde{g} : (\underline{x}, \bar{x}) \to (0, 1)$ with inverse $\tilde{g}^{-1} : (0, 1) \to (\underline{x}, \bar{x})$. Note that the interval $(\underline{x}, \bar{x})$ may be equal to $(-\infty, \infty)$. Examples of such a function g are the logistic, the piecewise-linear and the scaled arctangent function.

Auxiliary variable Gibbs sampling is possible if the density kernel p can be decomposed as follows:

$$p(x) \propto \pi(x) \prod_{k=1}^{K} l_k(x), \tag{2.63}$$

where π is a density kernel from which sampling is easy, and l_k $(k = 1, \ldots, K)$ are non-negative functions of $x \in \mathbb{R}^n$. The trick is that a set $U = (U_1, \ldots, U_K)$ of auxiliary variables is introduced such that a kernel of the joint density of X and U is given by:

$$p(x, u) \propto \pi(x) \prod_{k=1}^{K} I\left\{0 < u_k < l_k(x)\right\}. \tag{2.64}$$

It is easily seen that (2.63) is a marginal density kernel corresponding to the joint density (2.64). Therefore one can sample $X \sim p(x)$ by sampling both X and U from (2.64) and forgetting U.

Kernels from the conditional distributions of X and U are easily obtained from the joint density kernel:

$$p(x|u) \propto \pi(x) I\left\{l_k(x) > u_k, k = 1, \ldots, K\right\} \tag{2.65}$$

$$p(u|x) \propto \prod_{k=1}^{K} I\left\{0 < u_k < l_k(x)\right\} \tag{2.66}$$

It follows from (2.65) and (2.66) that an iteration of the auxiliary variable Gibbs sampler consists of drawing X from a truncated version of an 'easy' distribution with density kernel π, and sampling U_k $(k = 1, \ldots, K)$ from K independent uniform distributions.

We rewrite (2.61) as:

$$p(x) \propto \prod_{i=1}^{n} I\left\{\underline{x}_i < x_i < \bar{x}_i\right\} \prod_{h=1}^{H} \exp\left(c_h\, g\left(\sum_{i=1}^{n} a_{hi} x_i + b_h\right)\right). \tag{2.67}$$

which has the shape of (2.63) with

$$\pi(x) = \prod_{i=1}^{n} I\left\{\underline{x}_i < x_i < \bar{x}_i\right\}, \tag{2.68}$$

$$l_h(x) = \exp\left(c_h\, g\left(\sum_{i=1}^{n} a_{hi} x_i + b_h\right)\right) \quad \text{for } h = 1, \ldots, H. \tag{2.69}$$

where $\pi(x)$ is the 'easy' density kernel of n independent variables X_i $(i = 1, \ldots, n)$ with distribution $U(\underline{x}_i, \bar{x}_i)$.

Drawing U conditionally on the values of X is straightforward. Combining (2.66) and (2.69), it follows that the elements U_h $(h = 1, \ldots, H)$ are drawn independently from the distributions:

$$U_h | X = x \;\sim\; U\left(0, \exp\left[c_h\, g\left(\sum_{i=1}^{n} a_{hi} x_i + b_h\right)\right]\right) \tag{2.70}$$

Drawing X conditionally on the values of U is a little harder. We choose to break up X and sample the elements X_i $(i = 1, \ldots, n)$ conditionally on the values of U and the set of all other elements X_{-i}. Combining (2.65), (2.68) and (2.69), we derive a density kernel of the conditional distribution of X_i given X_{-i} and U:

$$p(x_i | u, x_{-i}) \propto I\left\{\underline{x}_i < x_i < \bar{x}_i\right\} I\left\{l_h(x_i, x_{-i}) > u_h, h = 1, \ldots, H\right\} \tag{2.71}$$

We now take a closer look at the inequalities $l_h(x_i, x_{-i}) > u_h$ $(h = 1, \ldots, H)$. First, we can rule out that $c_h = 0$ or $a_{hi} = 0$ for any h, since in that case we just delete the involved hidden neuron. If we consider $l_h(x_i, x_{-i})$ as a function of x_i for given values of x_{-i}, denoted by $l_{h,x_{-i}}(x_i)$, then the inverse $l^{-1}_{h,x_{-i}}$ (if it exists) is given by:

$$l^{-1}_{h,x_{-i}}(u_h) = \frac{1}{a_{hi}}\left(\tilde{g}^{-1}\left(\frac{\log(u_h)}{c_h}\right) - \left(\sum_{j=1, j\neq i}^{n} a_{hj} x_j + b_h\right)\right). \tag{2.72}$$

Note that this inverse exists only if $\log(u_h)/c_h \in (0, 1)$, and that the cases in which the inverse $l^{-1}_{h,x_{-i}}$ does not exist are the cases in which hidden neuron h implies no restriction

for x_i. Also notice that this implies an upper bound for x_i if $c_h a_{hi} > 0$ and a lower bound if $c_h a_{hi} < 0$.

We conclude that (2.71) is a density kernel of the distribution

$$X_i|U = u, X_{-i} = x_{-i} \ \sim \ U(x_{i,LB}(u, x_{-i}), x_{i,UB}(u, x_{-i})), \tag{2.73}$$

with

$$x_{i,LB}(u, x_{-i}) = \max\left\{ \max_{1\le h\le H} \left\{ l^{-1}_{h,x_{-i}}(u_h) \,\middle|\, c_h a_{hi} > 0, \frac{\log(u_h)}{c_h} \in (0,1) \right\}, \underline{x}_i \right\}$$

$$x_{i,UB}(u, x_{-i}) = \min\left\{ \min_{1\le h\le H} \left\{ l^{-1}_{h,x_{-i}}(u_h) \,\middle|\, c_h a_{hi} < 0, \frac{\log(u_h)}{c_h} \in (0,1) \right\}, \bar{x}_i \right\},$$

where $l^{-1}_{h,x_{-i}}(u_h)$ is given by (2.72), and where $[\underline{x}_i, \bar{x}_i]$ is the interval to which X_i ($i = 1, \ldots, n$) is a priori restricted.

The auxiliary variable Gibbs sampling procedure is now given by:

Initialization: Choose feasible $x^0 = (x_1^0, \ldots, x_n^0)$.

Do for $j = 1, 2, \ldots, m$

Do for $h = 1, 2, \ldots, H$

Obtain $u_h^j \sim U_h|X = x^{j-1}$ from (2.70).

Do for $i = 1, 2, \ldots, n$

Obtain $x_i^j \sim X_i|U = u^j, X_{-i} = x_{-i}^{j-1}$ from (2.73).

Here x_{-i}^{j-1} denotes

$$x_{-i}^{j-1} = x_1^j, \ldots, x_{i-1}^j, x_{i+1}^{j-1}, \ldots, x_n^{j-1},$$

the set of all components except x_i at their current values. Note that this procedure only requires drawing from uniform distributions, which is done easily and fast.

Chapter 3

Neural network sampling methods: improvements and strategies

Chapter 3 is based on Hoogerheide and Van Dijk (2006a).

3.1 Introduction

In chapter 2 a class of neural network sampling methods was proposed. It was shown that in an illustrative example the AdMit method, in which an Adaptive Mixture of t distributions is used as a candidate distribution, performed best among the neural network procedures. After that an example was given of a distribution for which the AdMit method outperformed competing methods, importance sampling and (both) the (independence chain and random walk) Metropolis-Hastings algorithm with a Student-t candidate and Gibbs sampling, where the latter got stuck in one of two far spaced modes for millions of draws. This chapter considers some changes that greatly improve the performance of the AdMit method and discusses the situations in which neural network sampling methods can be useful in general.

Section 3.2 discusses some improvements in both the construction and sampling parts of the AdMit procedure; the changes make the algorithm both faster and more reliable (in the sense of a quicker detection of distant modes). The effects of the improvements are illustrated in the example of a bimodal 3-dimensional posterior in an IV model for simulated data, previously considered in section 2.5. In section 3.3 the improved AdMit methods are applied to a 4-dimensional posterior distribution in a static 2-regime mixture model for US real GNP growth rates. The performance of the AdMit methods is compared

with two Gibbs sampling approaches, Gibbs sampling with data augmentation and the griddy Gibbs sampler. In section 3.4 it is illustrated that neural network sampling methods can especially be useful if one desires estimators of posterior characteristics with high precision. Secton 3.5 contains concluding remarks, among which some suggestions to extend the AdMit procedure.

3.2 Neural network sampling: some improvements

In section 2.5 a 3-dimensional example of a bimodal posterior target distribution in an instrumental variable (IV) regression model for simulated data was considered. The computing times (on an AMD Athlon^{TM} 1.4 GHz processor) of importance sampling (IS) and the (independence chain) Metropolis-Hastings (MH) algorithm with Type 3 neural network candidate density were reported in Table 2.5, together with the computing times of IS and independence chain MH with a unimodal candidate and a random walk chain MH algorithm. These Adaptive Mixture of t procedures (AdMit-IS and AdMit-MH) outperformed the competing methods in the sense that they gave estimates of posterior means with less variation (given the same amount of computing time).

In this section it is discussed how much computing time is required by the individual steps of the AdMit-IS method, i.e. both the steps of the construction of an approximation, a mixture of H Student-t densities $q_H(\theta)$, to the target density (kernel) $\tilde{p}(\theta)$ and the sampling method, where θ is a $k \times 1$ vector containing the parameters of interest of which one desires to evaluate properties such as the (posterior) mean and variance. This information is used to improve the AdMit procedure in such a way that it becomes both faster and more reliable in the sense of a higher probability that convergence to a 'proper' candidate density has been achieved at the end of the construction procedure. Note that when the AdMit procedure is applied to a multimodal target with huge distances between the modes, there is a chance that the resulting candidate will 'miss' certain modes, for example as the algorithm that is used to optimize the weight function (upon which the location of a new component in the candidate mixture is based) may get stuck in a (globally suboptimal) local optimum. The improvements discussed in this section cause a decrease of the chance that modes are 'missed' by the candidate density.

Table 3.1: Computing times of steps in Adaptive Mixture of t densities (AdMit) construction procedure

		computing time	
	step	procedure of chapter 2	improved procedure
0	initialization	0.1 s	0.7 s
1	evaluate (quality of) IS weights	11.0 s	74.6 s
2	determine mode μ_h, scale Σ_h of new component	0.6 s	50.3 s
3	optimize probabilities p_h of components	585.7 s	34.5 s
4	draw points from candidate mixture	0.9 s	23.3 s
	total	598.3 s	183.4 s

3.2.1 Improvements in the construction of an approximation to the target density

Table 3.1 shows how the computing time required for the construction of a (mixture of t) approximation to the non-elliptical 3-dimensional target density from section 2.5 is divided between the individual steps of the procedure. These reported computing times are the sums of the computing times required in the 15 iterations (yielding a mixture of 15 Student-t components). Note that a huge amount of time (98% of all computing time) is required for optimizing the probabilities of the components in the candidate mixture, the p_h $(h = 1, \ldots, H)$, where H takes the values $H = 2, \ldots, 15$ in the consecutive iterations of the algorithm. The reason is that this concerns the optimization of a non-linear function of p_h $(h = 1, \ldots, H)$, $E[w(\theta)^2]/E[w(\theta)]^2$ where $E[w(\theta)]$ and $E[w(\theta)^2]$ are given by:

$$E[w(\theta)^k] = \frac{1}{N} \sum_{i=1}^{N} \sum_{h=1}^{H} p_h \, w\big(\theta_h^i\big)^k \quad (k = 1, 2), \qquad w\big(\theta_h^i\big) = \frac{\tilde{p}(\theta_h^i)}{\sum_{l=1}^{H} p_l \, t(\theta_h^i | \mu_l, \Sigma_l, \nu)}. \tag{3.1}$$

Evaluating the function $E[w(\theta)^2]/E[w(\theta)]^2$ itself already requires $N\,H$ evaluations of the target density (kernel) $\tilde{p}(\theta_h^i)$ $(i = 1, \ldots, N; h = 1, \ldots, H)$ and $N\,H^2$ evaluations of the t density $t(\theta_h^i | \mu_l, \Sigma_l, \nu)$ $(i = 1, \ldots, N; h, l = 1, \ldots, H)$; the computation of (analytically evaluated) derivatives of $E[w(\theta)^2]/E[w(\theta)]^2$ with respect to p_h $(h = 1, \ldots, H)$ takes even more time.

One way to reduce the amount of computing time required for the construction of a (mixture of t) approximation, is to use different numbers of draws in different steps. One

can use a relatively small sample of N_1 draws for the optimization of the p_h's, and a large sample of N_2 draws in order to evaluate the quality of the current candidate mixture at each iteration (in the sense of the coefficient of variation of the corresponding IS weights) and in order to obtain an initial value for the algorithm that is used to optimize the weight function (that yields the mode of a new Student-t component in the mixture).

Notice that if one would simply use a small sample of N draws in each step, the procedure would have huge difficulties to detect distant modes in the case of multi-modality, and the evaluation of the quality of the candidate mixture would be less reliable. In that case it is quite uncertain whether the procedure has converged to a candidate density that is 'close' to the target density, when the algorithm stops. Therefore we only reduce the sample used in order to optimize the p_h's, while augmenting the sample used in the other steps.

Note that it is not necessary to find the globally optimal values of the p_h's; a 'good' approximation to the target density is all that is required. Furthermore, if the p_h's would result in a poor candidate, then this will show up in the next steps in which a large sample of N_2 points is drawn from the candidate mixture and the variation of the corresponding IS weights is investigated.[1] Moreover, the use of a large number of draws is especially important to prevent the algorithm from stopping too early (yielding either a very bad candidate density or a candidate that could have been hugely improved by taking a few more steps of the algorithm, adding a few components to the candidate mixture) and to detect distant modes in the case of a multi-modal target density. Therefore, not only the speed but also the reliability of the procedure is improved when using different numbers of draws in different steps in a clever way.

Another change that improves the algorithm's ability to detect distant modes of a multi-modal target density $\tilde{p}(\theta)$ is to try several initial values (instead of one initial value, the draw from the candidate mixture $q_{H-1}(\theta)$ obtained in the previous iteration with the highest IS weight) for the algorithm that optimizes the weight function, and – if the optimization algorithm yields different optima for different initial values – to use the point corresponding to the highest value of the weight function among the optima as the mode μ_H of the new component in the candidate mixture. We suggest the following two extra initial values: the draw with the highest IS weight in a set of N_2 drawings from a

[1] In order to make the stopping criterion stricter in the sense of a lower probability of stopping the method too early, the percentage of 10% (of relative change in the coefficient of variation of the IS weights that is required to make the algorithm proceed in the sense of adding another component) can be changed to lower values, e.g. 5%.

unimodal Student-t candidate $q_{Student,l}(\theta)$ $(l = 1, 2)$ with 1 degree of freedom with mode & scale equal to the estimated mean & covariance of the target, estimated either (1) using the latest candidate mixture $q_{H-1}(\theta)$, or (2) using the unimodal Student-t density $q_{Student,1}(\theta)$ that is used in order to find initial value (1).[2] The idea behind these initial values is that the scale of the candidate mixture $q_{H-1}(\theta)$ may be too small (as compared with the target density). Sequentially adapting the mode and scale of a unimodal Cauchy candidate density may enable one to properly augment the scale and thereby find (possibly distant) regions with target probability mass that where not yet 'covered' by candidate mixture.

The effects of the improvements that are described above on the results of the AdMit method in the 3-dimensional example of section 2.5 are as follows. In section 2.5 we used $N = 5000$ draws in each step. We now decide to use $N_1 = 1000$ draws for optimizing the p_h $(h = 1, \ldots, H)$, and $N_2 = 100000$ draws in the other steps. The resulting computing times are in the last column of table 3.1. First of all it should be noted that this improved procedure constructs a mixture of less components (13 instead of 15), explaining partly the smaller amount of required computing time. This mixture of 13 Student-t densities provides a somewhat better approximation to the target density in the sense of a smaller coefficient of variation of a sample of 100000 IS weights (1.37 instead of 1.47). The reason for the smaller number of required components (and the slightly better quality of the approximation) is that the modes of the components are better located. Figure 3.1 shows the locations of the modes μ_h $(h = 1, \ldots, H)$ in the $\pi_1 \times \beta$ space. The AdMit method of chapter 2 locates the first 8 modes $\mu_1, \mu_2, \ldots, \mu_8$ all in the same (bottom-right) ridge, whereas the improved procedure immediately finds the other (top-left) region of relatively high posterior density values. We conclude that the improved method is clearly more reliable in the sense of a lower probability of stopping the construction process at an iteration when not all regions with relative much target probability mass are 'covered' by probability mass of the candidate mixture.

The smaller number of iterations is certainly not the only reason why the improved AdMit method has become faster; less computing time is required per iteration. The use

[2]Of course, the unimodal Student-t densities $q_{Student,l}(\theta)$ $(l = 1, 2)$ may be very different from the candidate mixture $q_{H-1}(\theta)$. Therefore, the point θ with highest IS weight $\tilde{p}(\theta)/q_{Student,l}(\theta)$ does not necessarily have a high IS weight $\tilde{p}(\theta)/q_{H-1}(\theta)$ when using the latest mixture $q_{H-1}(\theta)$ as a candidate. However, this implies no problem, as this point is only used as an 'extra' initial value in the optimization algorithm (applied to the weight function $\tilde{p}(.)/q_{H-1}(.)$) and only the best of the outcomes of the optimization algorithm will be used.

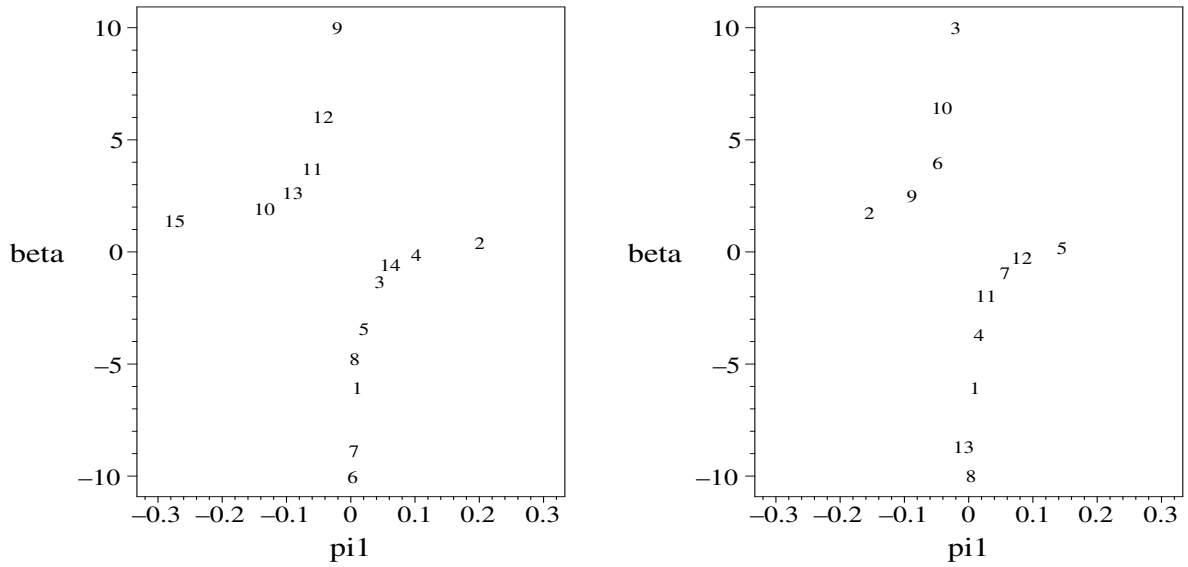

Figure 3.1: Locations of modes μ_h $(h = 1, 2, \ldots, H)$ in the $\pi_1 \times \beta$ plane of components of the candidate mixture in the AdMit procedure of chapter 2 (left) and the improved version (right)

of $N_2 = 1000$ instead of $N = 5000$ draws has caused a huge decrease of the computing time of step 3, the optimization of the p_h's. On the other hand the use of $N_2 = 100000$ instead of $N = 5000$ draws (and the use of several initial values in the algorithm for optimization of the weight function $\tilde{p}(.)/q_{H-1}(.)$) has resulted in only relatively small increases of the computing time of the other steps.[3] The AdMit construction method has become $598.3/183.4 = 3.26$ times faster.

[3]Note that sampling from the candidate mixture requires only very little time in case the same number of draws is used in each step, as in this case one can use the draws from the Student-t components that have already been generated in order to optimize the variation of the IS weights with respect to the p_h $(h = 1, \ldots, H)$. Note that such 'recycling' of draws during the construction of an approximation to the target density does not imply problems for the convergence of the AdMit-IS/MH estimates, as a *new* set of draws from a *fixed* candidate density is used to base the eventual estimates upon. The only thing that matters for the AdMit candidate construction phase is that it results in a candidate that is 'close' to the target. (The AdMit-MH method is *not* an 'adaptive Markov chain Monte Carlo' method for which the transition probabilities change during the sampling process.)

3.2.2 Improvements in the sampling procedure

Table 3.2 shows how the computing time required for importance sampling with target density from section 2.5 using $n = 1000000$ draws from a mixture of t candidate density is divided between the individual steps of the procedure. Note that a large amount of time (70% of all computing time) is required for evaluating the candidate density. The reason is that the AdMit program only required routines for the target density $\tilde{p}(\theta)$ and the basis density, the density of the mixture components, as a function of a single θ to be programmed, thereby allowing for an easier switch to other basis densities. However, for the Student-t density $t(\theta|\mu_h, \Sigma_h, \nu)$ one can easily program a routine that quickly transforms a matrix of rows θ_i' $(i = 1, \ldots, n)$ into a vector of elements $t(\theta_i|\mu_h, \Sigma_h, \nu)$ $(i = 1, \ldots, n)$.[4] The effect of this improvement on the AdMit method in the 3-dimensional example of section 2.5 is as follows. The resulting computing times are in the last column of table 3.2. First of all recall that the improved construction procedure resulted in a mixture of less components (13 instead of 15), explaining partly the smaller amount of required computing time for the evaluation of the candidate density. But the smaller number of components is certainly not the only reason why the improved AdMit method has become faster; less computing time is required per component. The AdMit importance sampling method has become $328.7/123.9 = 2.65$ times faster. The improvements discussed in this section have caused the total AdMit-IS procedure, consisting of both the construction of a candidate mixture and sampling from it, to become $(598.3+328.7)/(183.4+123.9) = 927.0/307.3 = 3.0$ times faster.

In the previous subsection it is mentioned that the changed AdMit method yields a somewhat better approximation to the target density than the method of chapter 2 in the sense of a smaller coefficient of variation of a sample of 100000 IS weights (1.37 instead of 1.47). Table 3.3 shows some sampling results for both methods. The numerical standard errors, relative numerical efficiencies (RNEs) and the influence of the 5% largest IS weights are all slightly better for the changed IS method, confirming the better quality of this changed AdMit-IS method. The acceptance rate and the serial correlation in the

[4]The sampling method that is used in order to obtain the results in chapter 2 required a 'for loop' over both n draws and H components, i.e. a 'for loop within a for loop'. However, using the matrix of rows θ_i' $(i = 1, \ldots, n)$ and a matrix of squares and cross products of the elements of θ_i' $(i = 1, \ldots, n)$ one can compute the vector of elements $t(\theta_i|\mu_h, \Sigma_h, \nu)$ $(i = 1, \ldots, n)$ by matrix operations, so that only one 'for loop' over the H components is needed.

Table 3.2: Computing times of steps in Adaptive Mixture of t densities (AdMit) sampling procedure (using 1000000 draws)

		computing time	
	step	procedure of chapter 2	improved procedure
1	sampling	36.9 s	30.6 s
2	evaluating target	13.2 s	13.2 s
3	evaluating candidate	231.7 s	33.7 s
4	evaluating IS weights	46.9 s	46.4 s
	total	328.7 s	123.9 s

MH Markov chain indicate that the changed method yields a somewhat beter candidate density in the (independence chain) MH algorithm.[5]

[5]The reported computing times for the MH algorithm are the same as for IS, whereas the MH algorithm requires an additional 'for loop'; however, the computing time for this is negligible as compared to the computing time of the rest of the method.

Table 3.3: Sampling results for the non-elliptically shaped posterior distribution in the IV regression (2.28) - (2.29) with $k = 2$ instruments for simulated data with $\pi = (0.1, 0.1)'$ (weak identification), $\sigma_{12} = 0.99$ (strong endogeneity)

		procedure of chapter 2		improved procedure	
	true values	AdMit IS	AdMit MH	AdMit IS	AdMit MH
$E(\pi_1)$	0.0199	0.0200	0.0195	0.0198	0.0199
(num. std. error)		($1.6\ 10^{-4}$)		($1.6\ 10^{-4}$)	
[RNE]		[0.3622]		[0.3651]	
$E(\pi_2)$	0.0157	0.0158	0.0153	0.0156	0.0157
(num. std. error)		($1.6\ 10^{-4}$)		($1.5\ 10^{-4}$)	
[RNE]		[0.3586]		[0.3707]	
$E(\beta)$	0.6404	0.6357	0.6531	0.6426	0.6415
(num. std. error)		(0.0065)		(0.0057)	
[RNE]		[0.2211]		[0.2893]	
$\sigma(\pi_1)$	0.0946	0.0945	0.0943	0.0945	0.0944
$\sigma(\pi_2)$	0.0935	0.0934	0.0934	0.0935	0.0934
$\sigma(\beta)$	3.0643	3.0745	3.0713	3.0622	3.0684
total time		927.0 s	927.0 s	307.3 s	307.3 s
time construction NN		598.3 s	598.3 s	183.4 s	183.4 s
time sampling		328.7 s	328.7 s	123.9 s	123.9 s
draws		$1\ 10^6$	$1\ 10^6$	$1\ 10^6$	$1\ 10^6$
time/draw		0.33 ms	0.33 ms	0.12 ms	0.12 ms
coeff. var. IS weights		1.47		1.37	
5% largest IS weights		27.3 %		25.1 %	
acceptance rate MH			32.5 %		34.4 %
serial corr. π_1			0.66		0.65
serial corr. π_2			0.66		0.65
serial corr. β			0.72		0.69

3.3 Example: mixture model for real US GNP growth

In this section the improved neural network sampling methods are applied to another non-elliptical target density $\tilde{p}(\theta)$, a 4-dimensional posterior density in a 2-regime mixture model for the US real GNP growth rate. The performance of the neural network sampling methods is compared with two Gibbs sampling methods, Gibbs sampling with data augmentation and the griddy Gibbs sampler, which do not get stuck in one of two far apart modes in this example.

Like in the example of the previous section, there are some bounds on the parameter values. However, in this section the bounds do not only exclude 'extreme' values (in the sense of relatively huge absolute values). Some of the bounds directly result from the nature of the parameters in the model. In the previous section the mode μ_H and scale Σ_H of the H-th component of the candidate mixture are obtained by the following steps. First, the weight function $\tilde{p}(\theta)/q_{H-1}(\theta)$ is optimized, where $q_{H-1}(\theta)$ is the candidate mixture obtained in the previous iteration of the algorithm. If the value $\theta_{max} \equiv \text{argmax}_\theta \, \tilde{p}(\theta)/q_{H-1}(\theta)$ does not occur at a boundary, then μ_H is given by θ_{max}, and Σ_H is minus the inverse of the Hessian of the logarithm of the weight function, evaluated at θ_{max}. If θ_{max} occurs at a boundary, μ_H and Σ_H are computed as the estimated mean and covariance matrix resulting from importance sampling with a 'residual' target density kernel $res(\theta) = \tilde{p}(\theta) - \tilde{c}\, q_{H-1}(\theta)$ and candidate $q_{H-1}(\theta)$, where $\tilde{c}$ is a positive constant. However, if the parameter bounds do not only exclude some 'extreme' values, but play a more important role in the model, one may expect that the maximum of the weight function $\tilde{p}(\theta)/q_{H-1}(\theta)$ will often be located at a boundary. Therefore, we choose to speed up the construction of the candidate mixture by skipping the optimization of the weight function $\tilde{p}(\theta)/q_{H-1}(\theta)$ and immediately computing μ_H and Σ_H as importance sampling results for a 'residual distribution'.[6]

In models for the growth rate of the real gross national product (GNP) one often allows for separate regimes for periods of recession and expansion. In this section a simple 2-regime model is considered, a static 2-regime mixture model. The data that are used are

[6]Note that this method in which the mode and scale matrix of components are based on importance sampling instead of optimization and the evaluation of a Hessian is not only a possibly useful alternative in the case of bounded domains; for example, if the evaluation of the Hessian of the logarithm of the weight function is difficult or very time consuming, this method may provide a large improvement in terms of required computing time.

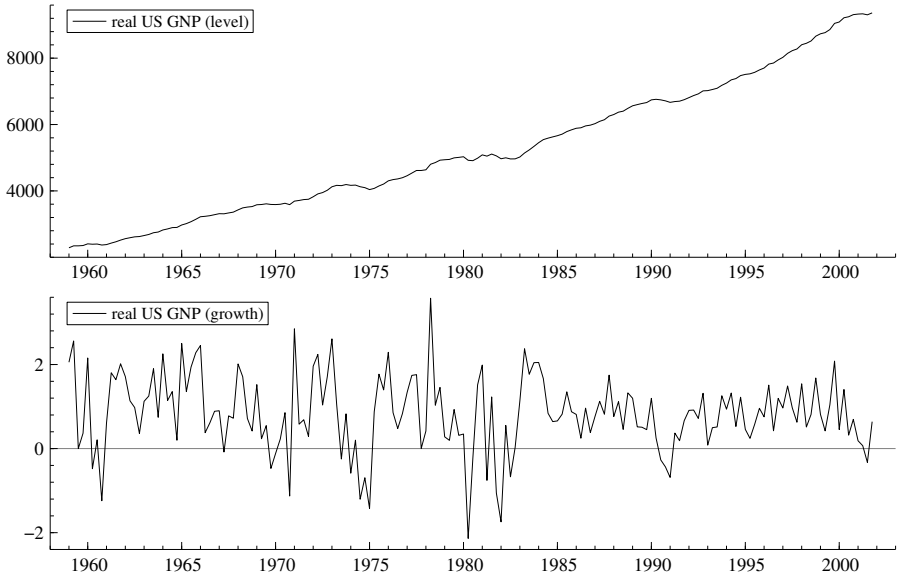

Figure 3.2: Real GNP of the USA in billions of dollars (above), and its quarterly growth rate in % (below). Source: Economagic.

the quarterly growth rates of the real US GNP in the period 1959-2001. The data are shown in Figure 3.2.

In this model the (percentage) growth rate y_t, defined as 100 times the first difference of the logarithm of real GNP, has two different mean levels:

$$y_t = \begin{cases} \beta_1 + \varepsilon_t & \text{with probability } p \\ \beta_2 + \varepsilon_t & \text{with probability } 1-p \end{cases}, \qquad t = 1, 2, \ldots, T, \tag{3.2}$$

where $\varepsilon_t \sim N(0, \sigma^2)$. For identification we assume that $\beta_1 < \beta_2$, so that β_1 and β_2 can be interpreted as the mean growth rates during recessions and expansions, respectively. For the parameters $(\beta_1, \beta_2, \sigma, p)$ the prior kernel is specified as $1/\sigma$ for $\beta_1 < \beta_2$, $0 \leq p \leq 1$, and 0 elsewhere. Furthermore β_1 and β_2 are restricted to the intervals $[-3, 2]$ and $[0, 3]$, respectively.

It should be noted that this model is merely used as an example to illustrate the (improved) neural network sampling methods in the case of a non-elliptical target distribution on a bounded domain, and to compare these with (Gibbs sampling with) data augmentation and the griddy Gibbs sampler. The assumption that the 'state' (recession/expansion) is independent over (quarterly) observations is obviously unrealistic.

We use both (improved) neural network algorithms AdMit-IS and AdMit-MH (with $N_1 = 1000$, $N_2 = 100000$) in order to obtain estimates of the posterior mean and standard deviation of β_1, β_2, σ and p. First, a candidate mixture of 5 Student-t distributions is

constructed. Figure 3.3 shows the shape of a highest posterior density (HPD) region of (β_1, β_2, p) conditional on $\sigma = 0.79$, the value of σ at the posterior mode (β_1, β_2, σ, p)[7], and the shape of a 'highest candidate density region' of (β_1, β_2, p) conditional on $\sigma = 0.79$. This illustrates once more that mixtures of t distributions can provide reasonable approximations to a wide variety of target distributions.[8] Table 3.4 shows the sampling results of AdMit-IS and AdMit-MH. For AdMit-MH the posterior mean and standard deviation of each parameter are estimated as the average and standard deviation of the draws; for AdMit-IS the weighted analogues are reported.

Figure 3.5 shows histograms, scaled so that these can be interpreted as estimates of marginal densities, of draws of β_1, β_2, σ, p obtained by the AdMit-MH method. Note the bimodality in the marginal posteriors of β_1 and p. The modes at $\beta_1 \approx -1$ and $p \approx 0.05$ correspond with the probability mass at the bottom of the left panel of Figure 3.3, whereas the modes at $\beta_1 \approx 0.7$ and $p \approx 1$ correspond with the probability mass at the top of the left panel of Figure 3.3. At the first region of the parameter space β_1 has the interpretation of the mean GNP growth rate during recessions which take place with low probability p. At the latter region of the parameter space β_1 has the interpretation of the mean GNP growth rate during periods of low or medium growth which occur with large probability p, while β_2 has the meaning of the mean GNP growth rate during 'exceptional expansions' (periods of very high growth rates) occuring with low probability $1 - p$. Notice that for $p = 0$ (or $p = 1$) the parameter β_1 (or β_2) is not identified. This local non-identification is reflected by Figure 3.3, in which for low values of p a wide spectrum of β_1 values is contained in the HPD credible set, whereas for high values of p the HPD region contain wide intervals of β_2 values. It is the latter situation that causes the modes at $\beta_1 \approx 0.7$ and $p \approx 1$ in the marginal posteriors.

It should be noted that one can also choose to identify β_1 and β_2 in another fashion than by imposing the restriction $\beta_1 < \beta_2$. One can instead impose the restrictions $\beta_1 < 0$ and $\beta_2 > 0$. In that case a less ill-behaved posterior results, as shown by Figure 3.4. However, this manner of identifying β_1 and β_2 is much more restrictive, as in that case only periods of negative (expected) growth are considered as recessions, whereas in practice periods of small, but positive growth rates are often also denoted as recessions. Further, although the posterior is less ill-behaved in that case, neural network sampling methods can still be

[7]The mode of the joint posterior of $(\beta_1, \beta_2, \sigma, p)$ is $(-1.00, 0.93, 0.79, 0.05)$.

[8]Note that the approximation is certainly not perfect; however, a better approximation requires (possibly much) more computing time in both the construction and sampling phase: there is a trade-off between the quality of the candidate mixture density and the speed of the construction and sampling.

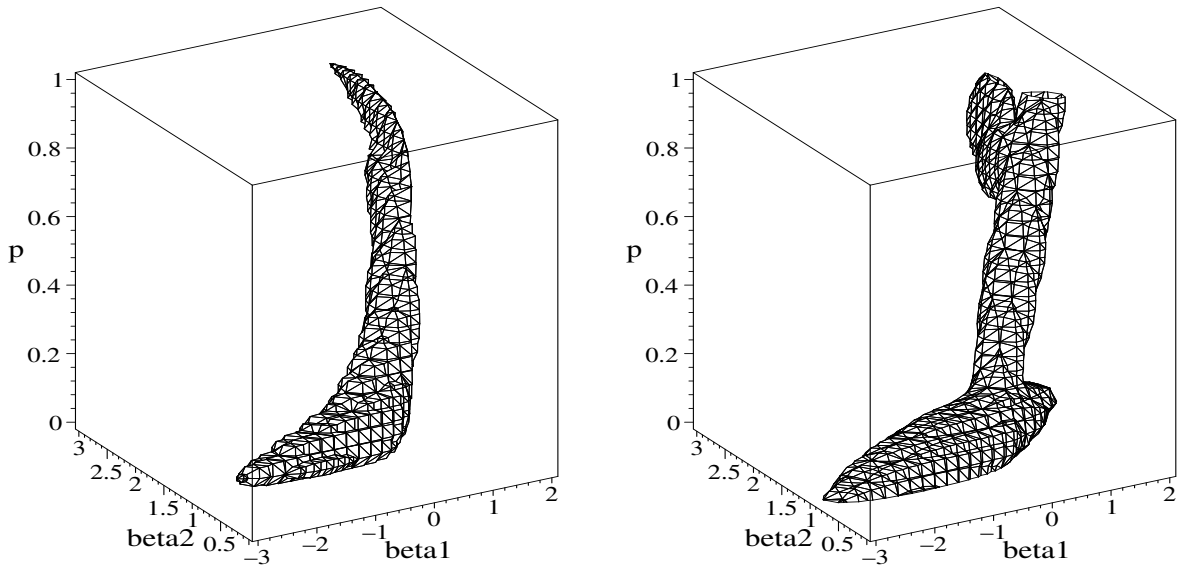

Figure 3.3: Highest posterior density (HPD) credible set (left) and 'highest candidate density' set (right) for a candidate mixture of 5 Student-t distributions for parameters (β_1, β_2, p) *in 2-regime mixture model for US real GNP growth rate (conditional on* $\sigma = 0.79$*, the value of* σ *at the posterior mode of* $(\beta_1, \beta_2, \sigma, p)$*)*

useful in cases of such posterior distributions that show somewhat non-elliptical shapes. A mixture of only 2 or 3 Student-t components, providing a good approximation to the posterior, can be quickly constructed, and importance sampling (or the Metropolis-Hastings algorithm) using this approximation as a candidate may outperform competing approaches. Finally, it should be noted that the main purpose of the 4-dimensional example in this section is to show the capabilities of neural network sampling methods in case of a highly non-elliptical posterior distribution.

In this model we can perform the method of Gibbs sampling with data augmentation of Tanner and Wong (1987). Data augmentation is used in order to sample from models with latent variables Z, in which directly sampling the parameters θ seems very difficult, but sampling θ given Z is straightforward. In this algorithm, the parameters θ are drawn conditionally on the latent variables Z, and the latent variables Z are drawn conditionally on θ. Forgetting the values of Z, this procedure yields a valid Markov chain for the parameters θ. In our model we define the latent variables Z_t $(t = 1, \ldots, T)$ as:

$$Z_t = \begin{cases} 1 & \text{if period } t \text{ is a recession period} \\ 0 & \text{if period } t \text{ is an expansion period} \end{cases} \qquad t = 1, 2, \ldots, T. \tag{3.3}$$

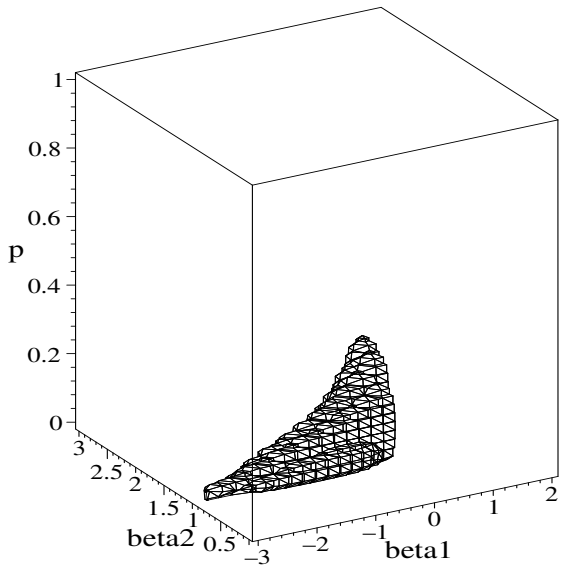

Figure 3.4: Highest posterior density (HPD) credible set for parameters (β_1, β_2, p) in 2-regime mixture model for US real GNP growth rate (conditional on $\sigma = 0.79$, the value of σ at the posterior mode of $(\beta_1, \beta_2, \sigma, p)$) under identification restrictions $\beta_1 < 0$, $\beta_2 > 0$ (instead of $\beta_1 < \beta_2$)

Conditionally on these latent variables Z (and each other), β_1 and β_2 are normally distributed, while σ^2 and p have an inverted gamma and a beta distribution, respectively. Conditionally on the values of the parameters, the latent variables Z_t $(t = 1, \ldots, T)$ have a Bernoulli distribution.[9]

Another Gibbs sampling approach that can be applied in this example is the griddy Gibbs sampling approach of Ritter and Tanner (1992). In this approach draws from the conditional distributions are obtained by applying the inversion method to a piecewise linear approximation to the conditional cumulative distribution function (CDF) that is computed using density (kernel) evaluations for a grid of input values.

[9]The distribution of (β_1, β_2) conditional on latent variables Z and on σ, p is given by $\beta_i \sim N(\bar{y}_i, \sigma/\sqrt{T_i})$ $(i = 1, 2)$ truncated to the region with $\beta_1 < \beta_2$, where $T_1 = \sum_{t=1}^{T} Z_t$ is the number of recession observations, $T_2 = T - T_1$ is the number of expansion observations, $\bar{y}_1$ and $\bar{y}_2$ are the average of the GNP growth rates y_t for recession $(Z_t = 1)$ and expansion $(Z_t = 0)$ observations, respectively. It may occur, and in fact does occur for this dataset, that either $T_1 = 0$ or $T_2 = 0$ during the data augmentation sampling process; if β_1 and β_2 would not be restricted to bounded intervals, the data augmentation procedure would 'crash'. And if these intervals for β_1 and β_2 would be chosen too wide, the procedure would get 'stuck' in a sequence of 'extreme' values for β_1 or β_2 (and $p = 0$ or $p = 1$).

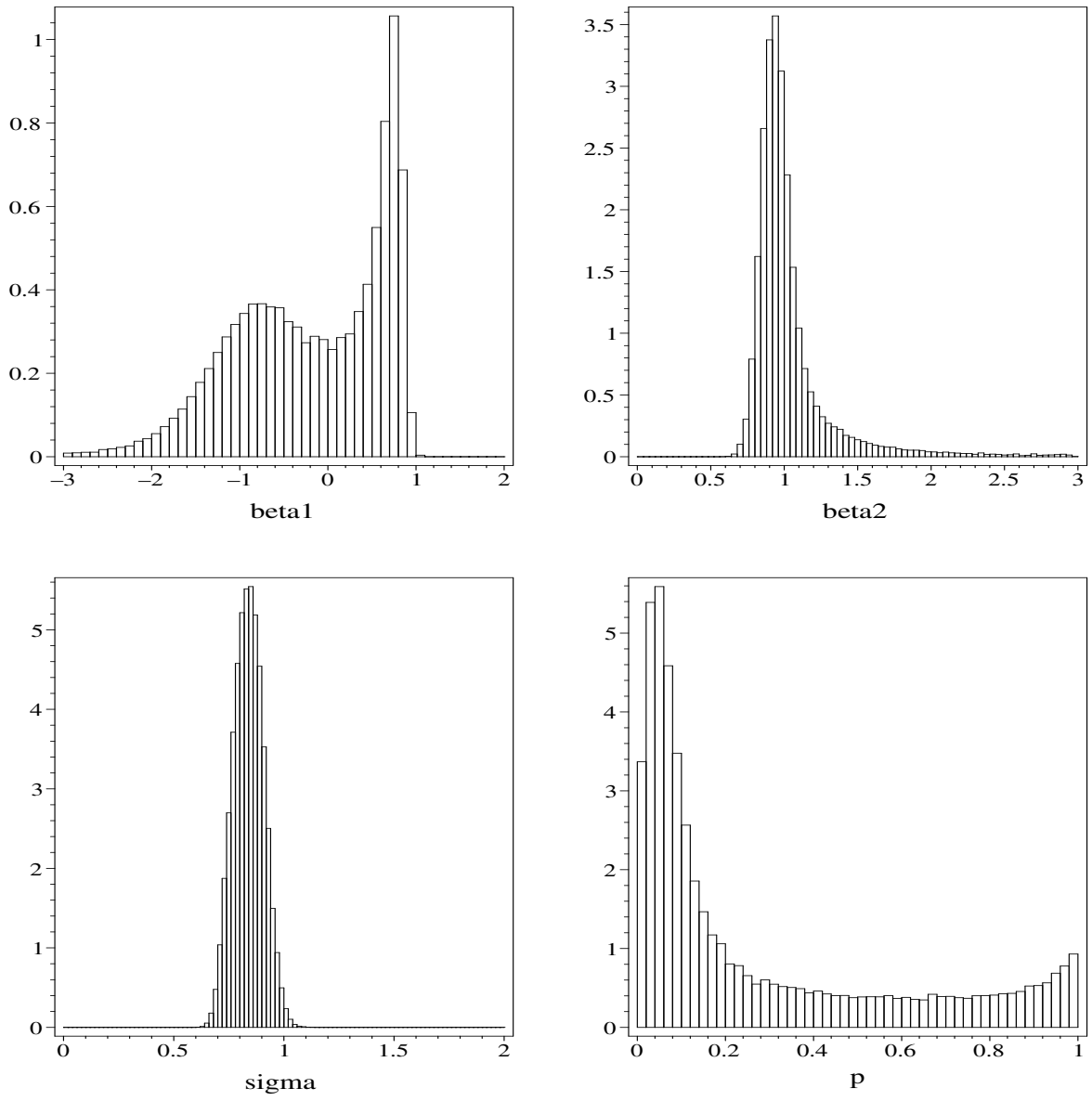

Figure 3.5: Histograms of draws of β_1, β_2, σ, p in AdMit-MH, scaled so that these can be interpreted as estimates of marginal densities

Table 3.4 shows sampling results for data augmentation and the griddy Gibbs sampler (that uses grids of 50 points for all four parameters). Again, the posterior mean and standard deviation of each parameter are estimated as the average and standard deviation of the draws.[10] The AdMit procedures beat the Gibbs samplers in the sense of yielding estimates with less variation in the same (or actually even somewhat less) computing time, where AdMit-IS outperforms AdMit-MH. Notice the huge serial correlation (especially the serial correlation of 0.993 for p) in the Gibbs sequence of the data augmentation method, which is even much higher than for the griddy Gibbs sampler: the extra elements Z_t $(t = 1, 2, \ldots, T)$ in the Gibbs sequence introduced by the data augmentation cause a large increase of the serial correlation. This huge serial correlation implies that the data augmentation estimates of the posterior means have a higher standard deviation (estimated by repeating the simulation 20 times) than the AdMit methods, even though AdMit-IS and AdMit-MH require the construction of a candidate mixture and the sampling takes somewhat more time per draw (0.14 ms versus 0.11 ms). Further notice that the evaluation of the target density over grids of points causes the griddy Gibbs sampler to be relatively very slow as compared to the AdMit methods and data augmentation: the griddy Gibbs sampler takes much more time per draw.

Finally, note that the data augmentation method requires more knowledge about the model in the sense of the specification of latent variables and derivation of conditional distributions, whereas the AdMit neural network methods (and the griddy Gibbs sampler) only require the evaluation of the posterior density kernel.

[10] It should be noted that the average of the draws is not the best estimate of the mean. It is better to use Rao-Blackwellization (*e.g.*, see Casella and George (1990) or Lancaster (2004)), which in this case amounts to using the average of the conditional means given draws of the other parameters. The average of the conditional means has less simulation noise than the average of the draws. Although Rao-Blackwellization is more natural in a Gibbs sampling approach, it is also possible to use the draws obtained in the AdMit-IS or AdMit-MH approach in order to compute Rao-Blackwellized estimates of the means and standard deviations. In principle, estimating the posterior mean of a parameter using Rao-Blackwellization is possible whenever one is able to obtain draws from the posterior distribution (of the other parameters) and knows the conditional posterior density of the parameter. In this section the average of the draws is used as an estimate for the mean, because of the simplicity of this approach, and because this method enables one to directly use the results of Geweke (1989) to compute numerical standard errors (and the corresponding relative numerical efficiency).

Table 3.4: Sampling results for the 2-regime mixture model

	Admit IS		Admit MH		Data Augmentation		Griddy Gibbs	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
β_1	-0.1795	0.8449	-0.1794	0.8427	-0.1681	0.8500	-0.1974	0.8491
(st.dev. 20×)	(0.0024)		(0.0038)		(0.0070)		(0.0260)	
(num. std. error)	(0.0022)							
[RNE]	[0.1431]							
β_2	1.0353	0.2841	1.0342	0.2817	1.0374	0.2849	1.0306	0.2754
(st.dev. 20×)	(0.0011)		(0.0017)		(0.0024)		(0.0071)	
(num. std. error)	(0.0012)							
[RNE]	[0.0570]							
σ	0.8388	0.0675	0.8390	0.0673	0.8395	0.0675	0.8369	0.0670
(st.dev. 20×)	(0.0002)		(0.0003)		(0.0002)		(0.0012)	
(num. std. error)	(0.0002)							
[RNE]	[0.1455]							
p	0.2788	0.3045	0.2777	0.3040	0.2856	0.3086	0.2729	0.3016
(st.dev. 20×)	(0.0009)		(0.0015)		(0.0038)		(0.0105)	
(num. std. error)	(0.0009)							
[RNE]	[0.1123]							
total time	204 s		204 s		264 s		251 s	
time construction NN	64 s		64 s					
time sampling	140 s		140 s		264 s		251 s	
draw	1000000		1000000		2500000		20000	
time/draw	0.14 ms		0.14 ms		0.11 ms		12.6 ms	
coeff. of var IS weights	2.55							
5% largest weights	44.2 %							
acceptance rate MH			21.1 %					
serial corr. β_1			0.79		0.90		0.76	
serial corr. β_2			0.83		0.80		0.63	
serial corr. σ			0.79		0.55		0.43	
serial corr. p			0.80		0.993		0.87	

3.4 When can neural network sampling methods be useful?

The examples of the 3-dimensional (highly non-elliptical) posterior distribution in the IV model and the 4-dimensional posterior in the mixture model for the US real GNP growth show estimators of posterior means with very small standard deviation, or equivalently very high precision (inverted variance). Neural network sampling methods yield these high precisions in a certain amount of time in which competing methods yield estimators with larger variance; in other words, competing methods require more computing time to reach the high precision resulting from the neural network approaches. However, if one

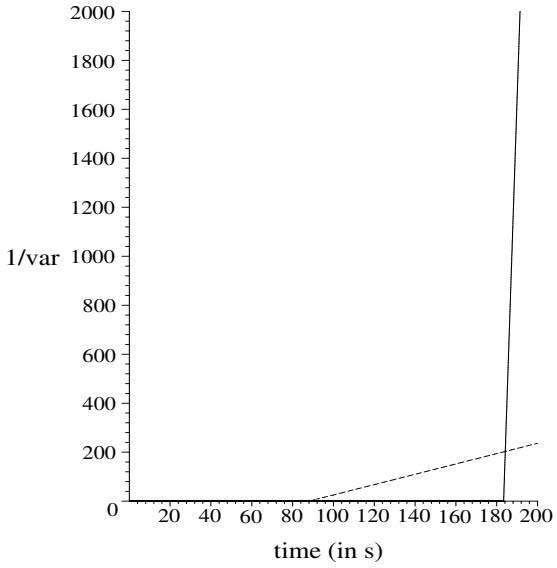

Figure 3.6: Precision (1/variance) of IS estimator of posterior mean of β for Student-t candidate density with scale and mode adapted to target density (dashed line) and for AdMit (mixture of t) candidate density (solid line) in the 3-dimensional bimodal posterior distribution in an IV model for simulated data, discussed in sections 2.5 and 3.2

would only require estimators of the posterior means with relatively low precision, then competing methods would outperform the neural network methods: IS with a unimodal Student-t candidate density (with mode and scale adapted to the target) in the IV model, and the data augmentation method in the mixture model. The reason for this is that the construction of a (mixture of t) approximation to the target distribution requires a certain amount of time; in (less than) this amount of time the IS approach with Student-t candidate or data augmentation has already started the sampling process and is therefore able to produce an estimator of the posterior mean (or other characteristics of interest), while the neural network approach is still in the 'candidate construction phase'. In other words, using an amount of time for the construction of a good candidate distribution is an 'investment' that is especially 'profitable' if one requires estimates with a very high precision, as the more draws are required from the candidate the relatively less 'expensive' is the investment in the construction of the candidate.

For the 3-dimensional non-elliptical posterior in the IV model this idea of the construction of a good candidate as an 'investment' is illustrated in Figure 3.6. Until 183.4 seconds the (improved) AdMit method is only constructing a candidate, while after 88

seconds (required for iteratively adapting the mode and scale to the target density) the IS approach with Student-t candidate is already sampling. Once the AdMit-IS method starts sampling, it soon outperforms IS with Student-t candidate: the lines cross at 184.2 s, at a precision of $1/\text{var}(\widehat{E(\beta)}) = 203.1$ (or at a standard deviation of $\text{st.dev}(\widehat{E(\beta)}) = 0.07$). AdMit-IS only requires 0.8 s to catch up with the 95 s of sampling of the IS with t candidate; the 'increase of precision per second of sampling' is about 120 times larger for AdMit-IS. The 'increase of precision per second of sampling' for the IS estimator of the mean of θ_j, the j-th element of θ, is given by:

$$\frac{\partial[1/\text{var}(\widehat{E(\theta_j)})]}{\partial t} = \frac{\#\text{draws per s}}{\text{var}(\theta_j)} \, \text{RNE}_{E(\theta_j)} \tag{3.4}$$

where the RNE (relative numerical efficiency) is the ratio between the (estimated) precision of the IS estimator of $E(\theta_j)$ and (an estimate of) the precision of an estimator of $E(\theta_j)$ based on direct sampling (with the same number of draws), see Geweke (1989). In the example of the non-elliptical posterior in the IV model the 'increase of precision per second of sampling' for the posterior mean of β is therefore given by

$$\frac{\partial[1/\text{var}(\widehat{E(\beta)})]}{\partial t} = \frac{1/(0.12 \cdot 10^{-3})}{(3.0622)^2} \, 0.2893 = 257.1$$

and

$$\frac{\partial[1/\text{var}(\widehat{E(\beta)})]}{\partial t} = \frac{1/(0.03 \cdot 10^{-3})}{(3.0622)^2} \, 0.0006 = 2.1$$

for AdMit-IS and IS with a Student-t candidate, respectively; note that we use the same (AdMit-IS) estimate of $\text{st.dev}(\beta)$ of 3.0622 in both formulas. So, if one desires to obtain an estimator of the posterior mean of β with standard deviation $\text{st.dev}(\widehat{E(\beta)})$ larger than 0.07, then IS with t candidate is a better choice than AdMit-IS in the sense of requiring less computing time; if one needs an estimator of the posterior mean of β with standard deviation $\text{st.dev}(\widehat{E(\beta)})$ smaller than 0.07, then AdMit-IS is the better choice as in this case AdMit-IS needs (possibly much) less computing time.

Whether neural network sampling methods can be useful does not only depend on the desired precision of the estimators of the characteristics of interest of the posterior distribution. This also depends on whether the posterior (target) distribution is (nearly) elliptical or (highly) non-elliptical. Neural network sampling methods outperformed competing methods in the two examples of highly non-elliptical target distributions in sections 2.5 and 3.3 with HPD credible sets in Figure 2.9 (consisting of two far apart parts) and

Figure 3.3 (consisting of one highly curved shape). In both examples the reason for the non-elliptical shape is local non-identification: for certain values of some of the parameters other parameters in the model are not identified. This suggests that neural network sampling methods can be especially useful in models with local non-identification. However, local non-identification does not always imply a (highly) non-elliptical posterior distribution, as the effect of certain tiny ridges or second modes with little probability mass may be neglectable. For example, section 2.5 also shows sampling results for (simulated data in) an IV model with strong instruments. In this case the neural network methods are slower than competing methods like Gibbs sampling or IS/MH with a normal or Student-t candidate distribution. However, often one does not know the shape of the posterior density in advance. Therefore it often seems a good idea to apply the following procedure. First, construct a neural network (mixture of t) approximation to the target distribution. Second, 'prune' the neural network in the sense of dropping components that hardly improve the quality of the candidate density, and optimize the degrees of freedom of the Student-t components of candidate mixture (possibly consisting of only one component). Third, use the resulting candidate density in IS or the MH algorithm. This way the neural network samping method only takes a fixed amount of extra computing time at the beginning, but is at least as good as IS (or the MH algorithm) using a normal or Student-t candidate during the sampling process.

Note that the possible usefulness of 'pruning' is not restricted to cases of (nearly) elliptical target distributions. In general, one can check the usefulness of each component in the candidate mixture between the construction and sampling phases. The choice of whether or not to drop a component can be based on the trade-off between the extra time it requires during the sampling process and the loss of quality of the candidate density (represented in formula (3.4) by the number of draws per second and the RNE, respectively).

Finally, if it is possible to divide target density of the parameters $\theta = (\theta^a, \theta^b)$ into a marginal target density of θ^a for which an explicit formula $\tilde{p}(\theta^a)$ is available and a conditional $\tilde{p}(\theta^b|\theta^a)$ that is easy to sample from, then it is obviously a good idea to apply a neural network sampling method to draw from the marginal density $\tilde{p}(\theta^a)$ and use the true conditional density $\tilde{p}(\theta^b|\theta^a)$ to sample θ^b. For example, if we would not restrict the parameters π in the IV model (with flat prior) to a certain bounded area, an explicit formula for the marginal density of β is known, while the conditional distribution of π given β is simply Student-t.

3.5 Concluding remarks

In chapter 2 a class of neural network sampling methods was proposed. It was shown that the AdMit method, in which a mixture of Student-t distributions is used as a candidate distribution, performed best among the neural network procedures, and an example was given of a distribution for which the AdMit method outperformed competing methods, importance sampling and (both) the (independence chain and random walk) Metropolis-Hastings algorithm with a Student-t candidate and Gibbs sampling, where the latter got stuck in one of two far spaced modes for millions of draws. This chapter discussed some large improvements in the AdMit method; these improvements make the method faster (about three times as fast in an example of a 3-dimensional bimodal target distribution) and more reliable (in the sense of a quicker detection of distant modes). The improved AdMit methods are applied to the 4-dimensional posterior distribution in a mixture model for US real GNP growth rates. The AdMit methods outperform two Gibbs sampling approaches, Gibbs sampling with data augmentation and the griddy Gibbs sampler; in this case the Gibbs sequences did not get stuck in one of two modes – in fact the joint posterior density is unimodal in this example – but the high serial correlation in the Gibbs sequences caused the Gibbs samplers to yield estimators of posterior moments with larger standard deviations than those resulting from the neural network methods (in the same computing time). Finally, it is illustrated that neural network sampling methods can especially be useful if one desires estimators of posterior characteristics with high precision.

We end this chapter with some remarks on how to extend the proposed methods. A straightforward alternative is to use a neural network function as a candidate density in rejection sampling instead of importance sampling or the Metropolis-Hastings algorithm. Another extension that is more difficult to implement, but much more interesting for practical purposes is to build a neural network method within a Gibbs sampling procedure (or a 'MH within Gibbs' algorithm). If it is hard to draw from one of the conditional distributions, say the conditional distribution of θ^a given θ^b with $\theta = (\theta^a, \theta^b)$ where θ^a and θ^b both consist of multiple elements, two options are to use a 'MH within Gibbs' step or to use several steps of the griddy Gibbs sampler. For a 'MH within Gibbs' step a candidate density is required. An option is to approximate the conditional target density of θ^a given θ^b with a mixture of Student-t densities. However, the disadvantage is that in each iteration (for each different value of θ^b) a new approximation has to be constructed, which can result in a very time consuming algorithm. In order to keep the computing

time for obtaining approximations to conditional target densities relatively small, one can store both the θ^b's and the approximations to the conditional densities of θ^a given θ^b. In each iteration one can use as an initial point the (mixture of t) approximation for the value of θ^b that is closest to the current value of θ^b (taking into account the scales and correlations of the elements of θ^b) among the set of previous θ^b's in the Markov chain. After that, one may add one or more components to the candidate mixture and drop (almost) useless components in order to prevent ending up with mixtures of huge numbers of components. Nevertheless, the resulting algorithm will still be rather slow. However, a 'MH withing Gibbs' step with a poor candidate distribution may result in a very low acceptance probability, resulting in very slow convergence of the estimators, or even an unreliable algorithm in which certain regions of the domain of θ that contain substantial probability mass may be 'missed'. And the use of several griddy Gibbs steps also yields a slow algorithm, in which the division of the sampling of θ^a into individual steps for sampling the elements of θ^a may seriously increase the serial correlation in the Gibbs sequence. Therefore, the combination of neural network sampling methods and the Gibbs sampler (or the 'MH within Gibbs' algorithm) is an interesting topic for further research.

Part II

Instrumental variables

Chapter 4

Bayesian analysis of the instrumental variables regression model under a flat, Jeffreys or hierarchical prior

Chapter 4 is based on Hoogerheide, Kaashoek and Van Dijk (2004, 2006).

4.1 Introduction

In this chapter we consider Bayesian analysis of instrumental variables regression models under three priors: the flat prior, the Jeffreys prior and a hierarchical prior proposed by Chamberlain and Imbens (1996). In section 4.2 we consider shapes of posterior densities and credible sets under the flat prior and Jeffreys prior for several levels of endogeneity and instrument strength. Some explanations for these shapes are given in the simple case of one endogenous explanatory variable and one instrument. In section 4.3 the behavior of the marginal posterior of the structural parameter of interest is investigated in the case of many irrelevant instruments: the performance of the hierarchical prior is compared with that of the flat and Jeffreys prior. Section 4.4 contains concluding remarks.

4.2 Shapes of posterior densities in instrumental variables regression model with several degrees of endogeneity and instrument quality

Consider the following possibly overidentified Instrumental Variables (IV) model, also known as the incomplete simultaneous equations model (INSEM). Following Hausman (1983), let:

$$y = x\beta + \varepsilon \tag{4.1}$$

$$x = Z\Pi + v \tag{4.2}$$

where y is a $(T \times 1)$ vector of observations on the endogenous variable that is to be explained, x is a $(T \times 1)$ vector of observations on the explanatory endogenous variable, Z is a $(T \times k)$ matrix of weakly exogenous variables; β is a scalar structural parameter of interest, Π is a $(k \times 1)$ vector of reduced form parameters.[1] The restricted reduced form corresponding to the structural form (4.1)-(4.2) is given by:

$$y = Z\Pi\beta + u \tag{4.3}$$

$$x = Z\Pi + v \tag{4.4}$$

where $u \equiv v\beta + \varepsilon$. The error terms in the structural form and the restricted reduced form have covariance matrix Σ and Ω, i.e. $(\varepsilon_i, v_i)' \sim N(0, \Sigma)$ and $(u_i, v_i)' \sim N(0, \Omega)$, with

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix} = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix} \Sigma \begin{pmatrix} 1 & 0 \\ \beta & 1 \end{pmatrix} = \begin{pmatrix} \sigma_{11} + 2\sigma_{12}\beta + \sigma_{22}\beta^2 & \sigma_{12} + \sigma_{22}\beta \\ \sigma_{12} + \sigma_{22}\beta & \sigma_{22} \end{pmatrix}.$$

A well-known example is the wage regression where y is the logarithm of wage and x denotes the number of years of education which is possibly endogenous owing to the omission of a variable measuring (unobservable) ability. The problem is that potential instruments for x are hard to find as these variables must be correlated with education but uncorrelated with unobserved ability. Angrist and Krueger (1991) suggest using quarter of birth to form instrumental variables since quarter of birth affects the age at school entry. This model will be considered in chapter 5.

[1] A different convention is to use the notation y_1, y_2, X instead of y, x, Z (see *e.g.* Zellner et al. (1988)).

In this section we consider the shapes of posteriors in the instrumental variables regression model for simulated data for different levels of endogeneity and instrument strength. Subsections 4.2.1 and 4.2.2 contain results for the flat and Jeffreys prior, respectively.

4.2.1 Posteriors and credible sets under flat prior

In this subsection we specify the following non-informative prior density of Drèze (1976):

$$p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-h/2} \text{ with } h > 0. \tag{4.5}$$

Given the model (4.1)-(4.5), one can easily derive the likelihood function and the posterior density kernel of (β, Π, Σ). Using properties of the inverted Wishart distribution (see Zellner (1971) and Bauwens (1990)) in order to integrate Σ out of the joint posterior, and choosing $h = 3$ in the prior density kernel (4.5) leads to the following joint posterior kernel of (β, Π):[2]

$$p(\beta, \Pi | y, x, Z) \propto \left| \begin{array}{cc} (y - x\beta)'(y - x\beta) & (y - x\beta)'(x - Z\Pi) \\ (x - Z\Pi)'(y - x\beta) & (x - Z\Pi)'(x - Z\Pi) \end{array} \right|^{-T/2}. \tag{4.6}$$

For $\Pi = 0$ the posterior kernel in (4.6) reduces to the (non-zero) constant $((y'y)(x'x) - (y'x)^2)^{-T/2}$, so that for $\Pi = 0$ the conditional posterior density of β is improper. For $\Pi \neq 0$ the integral $\int p(\beta, \Pi | y, x, Z) d\beta$ is finite; however, when $\Pi \to 0$ this integral increases at a rate of $(\Pi' Z' M_x Z \Pi)^{-1/2}$, so for the just identified case of $k = 1$ the integral $\iint p(\beta, \Pi | y, x, Z) d\beta d\Pi$ is not finite. For details, see Propositions 3 and 4 below. So, in the exact identified case the joint density of β and Π is improper on $\mathbb{R}^{k+1}$. Although improper on $\mathbb{R}^{k+1}$, the posterior in (4.6) can be made proper by restricting β and/or Π to a certain area. In that case it depends on the data y, x and Z, whether the behavior for $\Pi = 0$ still dominates the analysis.

For illustrative purposes, the posterior kernel in (4.6) is calculated for simulated data sets from (4.1) - (4.2) with $k = 1$, $T = 100$, $\beta = 0$, $\sigma_{11} = \sigma_{22} = 1$ for nine cases. In each case we use the same vector of instruments denoted by z, where the elements of z are

[2]This prior with $h = 3$ slightly differs from the prior used by Drèze (1976) which would have $h = 4$ in this case. A reason for the choice of $h = 3$ is that the specification with $h = 3$ leads to a marginal posterior of (β, Π) that is equal to the concentrated likelihood function for (β, Π), so that the shapes of the marginal posterior of (β, Π) can immediately also be interpreted as the shapes of this concentrated likelihood function. Further, it should be noted that in this case with $T = 100$ data the differences in shapes between $h = 4$ and $h = 3$ are only minor, as this merely amounts to a difference between $-(T + 1)/2$ and $-T/2$ in the exponent of (4.6).

i.i.d. N(0,1) draws. Three different cases of identification (or quality of instruments) are considered: non identification/irrelevant instruments ($\Pi = 0$); weak identification/weak instruments ($\Pi = 0.1$); strong identification/strong instruments ($\Pi = 1$). These cases are combined with three cases of endogeneity, i.e. three different values of the correlation $\rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ between the error terms ε and v: strong ($\rho = 0.99$), medium ($\rho = 0.5$) and no ($\rho = 0$) degree of endogeneity. Figure 4.1 shows contour plots of the joint posterior kernel of β and Π in (4.6) for our nine simulated data sets; the posterior kernels are normalized over the displayed range. The contour plots reveal that there are three typical shapes of the graph of the joint posterior of β and Π: bell-shape, bimodality and elongated ridges.[3] Table 4.1 gives an overview of the possible shapes of the joint posterior kernel of β and Π in this simple IV regression model with $k = 1$ instrument for different cases of simulated data.[4]

Note that in the three cases of simulated data sets with strong instruments ($\Pi = 1$), the contour plots do not show a high-level ridge at $\Pi = 0$; the value of the joint posterior kernel for $\Pi = 0$ is relatively very small as compared to the value of the joint posterior kernel at its mode. In the just identified model (with $k = 1$) the mode is given by $(\hat{\beta}_{2SLS} = y'z/x'z, \hat{\Pi}_{OLS} = x'z/z'z)$, and the ratio between the posterior kernel in (4.6) for $\Pi = 0$ (and arbitrary β) and the value at its mode $(\hat{\beta}_{2SLS}, \hat{\Pi}_{OLS})$ is:

$$\frac{p(\beta, \Pi = 0|y, x, Z)}{p(\hat{\beta}_{2SLS}, \hat{\Pi}_{OLS}|y, x, Z)} = \left[1 - \frac{r_{x,z}^2 + r_{y,z}^2 - 2r_{y,z}r_{x,z}r_{y,x}}{1 - r_{y,x}^2}\right]^{T/2}, \tag{4.7}$$

where $r_{x,z} \equiv x'z/\sqrt{x'x\ z'z}$, etc. In the three cases of strong instruments (with large $r_{x,z}^2$) as well as in the case of weak instruments and strong endogeneity (with $r_{y,x}^2$ close to one) the ratio (4.7) is small ($< 10^{-9}$). Also in the case of irrelevant instruments and strong endogeneity $1 - r_{y,x}^2$ is small; however, as $r_{y,z}$ and $r_{x,z}$ are both small and $r_{y,z} \approx r_{x,z}$, the numerator on the right-hand side of (4.7) is even much smaller, so that in this case the contour plot displays a high-level ridge at $\Pi = 0$.

If we consider the contour plot of the posterior kernel (4.6) raised to the power 1/20, so that the contour plot also shows the contours for much lower values of the posterior

[3]The same types of shapes of posterior distributions also occur in the vector error correction model (VECM) under cointegration, another reduced rank regression model. The VECM under cointegration is mathematically equivalent to the IV regression model. A comparison of classical tests in both models is provided by Hoogerheide and Van Dijk (2001).

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

kernel, we observe also in the case of strong identification the presence of bimodality or an elongated ridge around the line $\Pi = 0$; see Figure 4.2. The origin of these hyperbolic contour lines becomes intuitively clear if we consider the fact that the structural form (4.1)-(4.2) is equivalent with the orthogonal structural form (see Zellner et al. (1988)):

$$y = x\beta + v\phi + \eta \tag{4.8}$$

$$x = Z\Pi + v \tag{4.9}$$

where $\phi = \sigma_{12}/\sigma_{22}$; η and v are mutually independent, i.i.d. Gaussian error terms. The equation (4.8) is equivalent with

$$y = x\gamma_1 + Z\gamma_2 + \eta \tag{4.10}$$

where $\gamma_1 = \beta + \phi$, $\gamma_2 = -\Pi\phi$, so that $\gamma_2 = \Pi(\beta - \gamma_1)$, and in the case of $k = 1$ instrument $\beta = \gamma_1 + \gamma_2/\Pi$. In the just identified model the set of points (β, Π) for which the posterior kernel in (4.6) scaled by the value at its mode, $p(\hat{\beta}_{2SLS}, \hat{\Pi}_{OLS}|y, x, Z)$, has a certain value $C \in (0, 1]$ is given by

$$\{\beta = \hat{\gamma}_1 + \hat{\gamma}_2/\Pi \pm \sqrt{p_C(\Pi)}/\Pi, p_C(\Pi) \geq 0, \Pi \neq 0\} \tag{4.11}$$

where $(\hat{\gamma}_1, \hat{\gamma}_2)$ are the OLS estimators in (4.10), and $p_C(\Pi)$ is a polynomial of degree 2 that is non-negative on the interval with bounds $\hat{\Pi}_{OLS} \pm \frac{s_x}{s_z}\sqrt{(1 - r_{x,z}^2)(C^{-2/T} - 1)}$ with $s_x \equiv \sqrt{x'x}$, which includes $\Pi = 0$ for C small enough; so there are hyperbolic contour lines in a neighborhood of $\Pi = 0$ for any level of endogeneity/instrument strength, although the level of endogeneity/instrument strength determines the relative level of the posterior around $\Pi = 0$. In the three cases of a strong instrument $\hat{\Pi}_{OLS}$ is far from zero (with t-value of $\hat{\Pi}_{OLS}$ larger than 10), resulting in (nearly) elliptical shapes far away from $\Pi = 0$. In the cases of no/weak identification $\hat{\Pi}_{OLS}$ is small (with t-value smaller than 1). In these cases the shapes depend on $\hat{\gamma}_2$: if (the t-value of) $\hat{\gamma}_2$ is close to zero, the contour plot shows connected ridges around $\Pi = 0$; otherwise it displays two disconnected ridges on both sides of $\Pi = 0$ (and on both sides of $\beta = \hat{\gamma}_1$ where $\hat{\gamma}_1 \approx 1$ for this simulated data set). The squared t-value of $\hat{\gamma}_2$, the F-statistic, is equal to

$$t_{\hat{\gamma}_2}^2 = (T - 2)\left(\frac{(1 - r_{y,x}^2)(1 - r_{x,z}^2)}{(r_{y,z} - r_{x,z}r_{y,x})^2} - 1\right)^{-1},$$

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

Table 4.1: Shape of the posterior density kernel of β and Π in the IV regression model (4.1)-(4.2) with one instrument and weak prior (4.5) for nine situations

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

then the expression $(1 - r_{y,x}^2)(1 - r_{x,z}^2)/(r_{y,z} - r_{x,z}r_{y,x})^2$, which is always larger than 1, is close to 1 so that $t_{\hat{\gamma}_2}^2$ is relatively large.

As can be seen from Figure 4.2 and formula (4.11), even in the presence of strong instruments and no/medium endogeneity the contours are, strictly speaking, not elliptical. However, if one restricts the region of integration to a certain bounded area the influence of these tiny ridges on inference is negligible; then one may for practical purposes consider the joint posterior distribution of β and Π as elliptical.

So, the posterior density kernel of β and Π may show highly non-elliptical shapes if instruments are weak. Drèze(1976, 1977) and Kleibergen and Van Dijk (1994b, 1998) present theoretical results on the conditional and marginal distributions of β and Π corresponding to this joint density kernel. We reformulate and illustrate their results for the simple IV regression model (4.1)-(4.5), and give explanations for some shapes of marginal distributions.

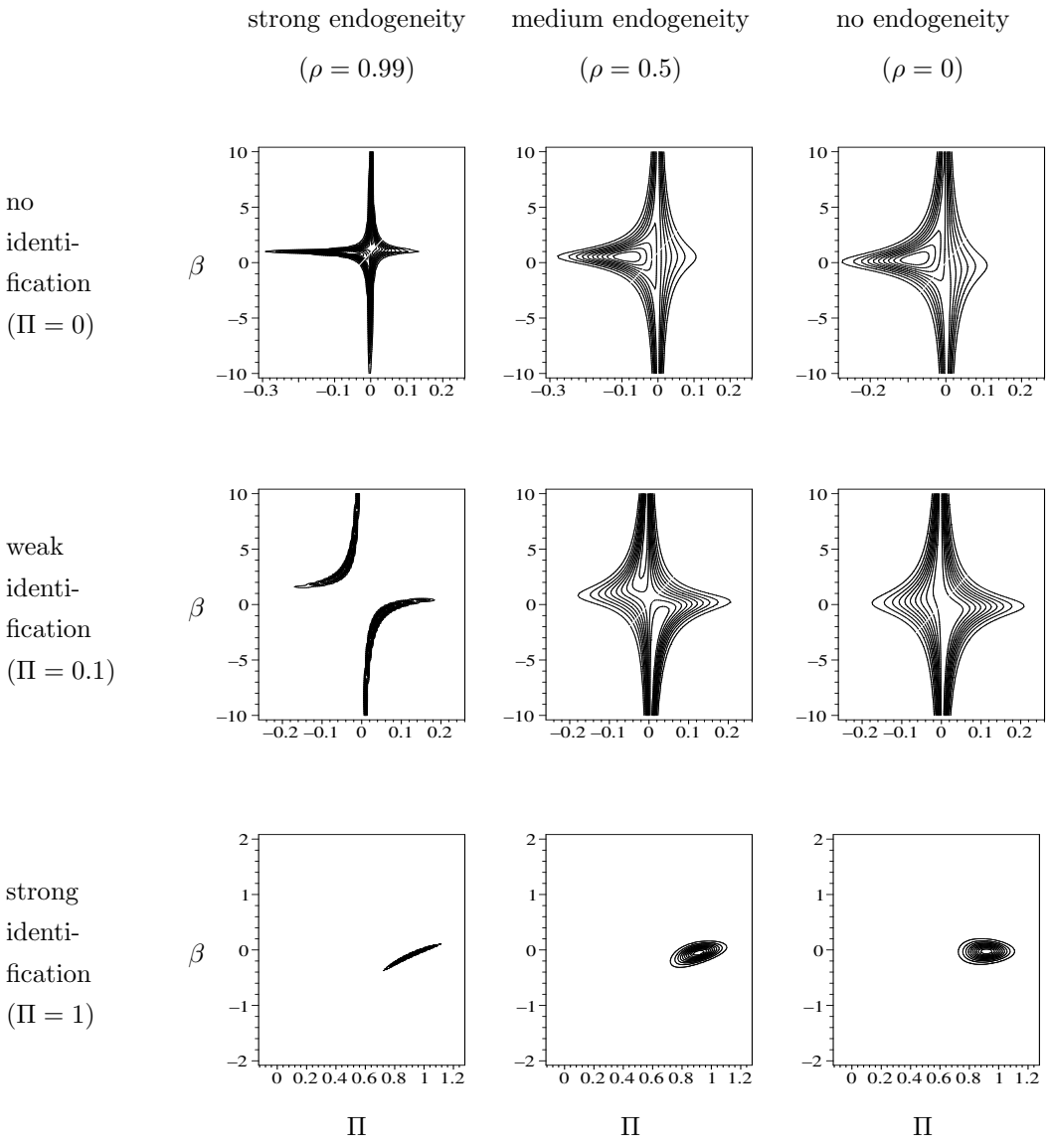

Figure 4.1: Contour plots in the $\Pi \times \beta$ plane: joint posterior kernel of Π and β in (4.6) in IV model under flat prior for nine simulated data sets; three cases of identification ($\Pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity)

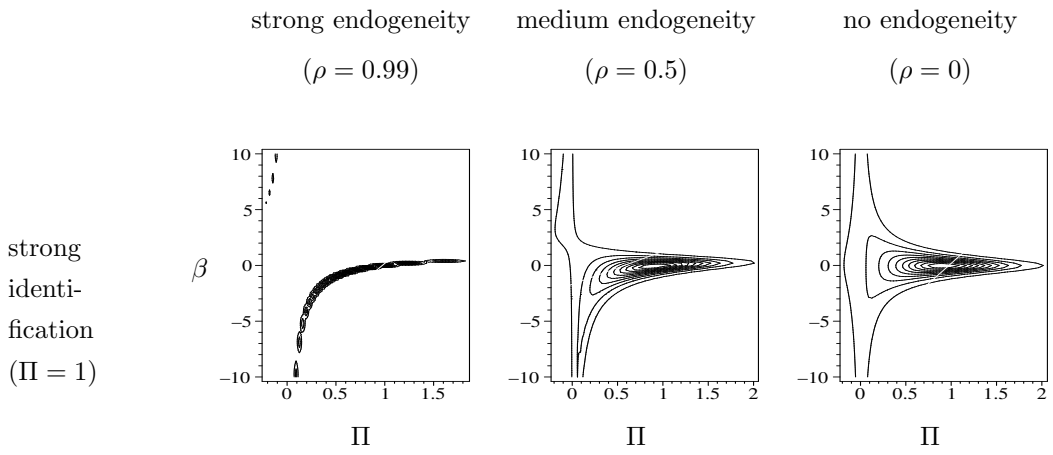

Figure 4.2: Contour plots in the $\Pi \times \beta$ *plane: joint posterior kernel of* β *and* Π *in (4.6) raised to the power 1/20 in IV model for three simulated data sets; the case of strong identification (*$\Pi = 1$*) combined with three levels of endogeneity (*$\rho = 0.99,\ 0.5,\ 0$ *corresponding to strong, medium, no endogeneity)*

Weak and strong structural inference

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

Proposition 1: Conditional posterior of β **given** Π**:** *In the IV regression model (4.1)-(4.2) with prior (4.5) the conditional posterior density of* β *given* Π *(with* $\Pi \neq 0$*) is a Student t density with mode* $\hat{\beta} \equiv (x'M_v x)^{-1}(x'M_v y)$*, scale* $s^2_{\hat{\beta}}(x'M_v x)^{-1}$ *and* $(T-1)$ *degrees of freedom:*

$$p(\beta|\Pi, y, x, Z) = \frac{c}{\sqrt{s^2_{\hat{\beta}}(x'M_v x)^{-1}}} \left[1 + \frac{1}{T-1}\frac{(\beta-\hat{\beta})^2}{s^2_{\hat{\beta}}(x'M_v x)^{-1}}\right]^{-T/2} \tag{4.12}$$

where $(T-1)s^2_{\hat{\beta}} \equiv (y - x\hat{\beta})'M_v(y - x\hat{\beta})$ *and* c *is a constant that only depends on* T*;* $M_v \equiv I - v(v'v)^{-1}v'$ *with* $v \equiv x - Z\Pi$*, i.e.* v *is a function of the parameter* Π *(and the data* x*,* Z*) instead of the vector of actual error terms.*

For $\Pi \to 0$ the conditional posterior variance of β tends to ∞ as in this case $x'M_v x \to 0$ (if $\Pi = 0$ then $v \equiv x - z\Pi = x$). For $\Pi = 0$ the conditional posterior density of β is improper. For $\Pi \neq 0$ conditional HPD credible sets of β are elliptical; in this case the conditional mean is equal to the OLS estimator of β in the orthogonal structural form

equation (4.8).

Proposition 2: Marginal posterior of β: *In the IV regression model (4.1)-(4.2) with prior (4.5) the marginal posterior density of β is proportional to the following ratio of two Student t kernels:*

$$p(\beta|y,x,Z) \propto \frac{[(y-x\beta)'(y-x\beta)]^{-(T-1)/2}}{[(y-x\beta)'M_Z(y-x\beta)]^{-(T-k-1)/2}}, \tag{4.13}$$

known as the 1-1 ratio or poly t density.

Structural inference on β depends on the level of identification. Moments exist up to the order of overidentification $(k-1)$. The marginal posterior of β tends to a bell-shaped function as long as the number of instruments k becomes large enough, which seems to be a paradoxical result: the presence of many (possibly *irrelevant*) instruments gives a bell-shaped function. In other words, even if the *quality* of the instruments is poor, a large *number* of instruments still yields a bell-shaped marginal posterior of β. This result appeared in an informal way in Maddala (1976), commenting on Drèze (1976). It should be noted that the location of the bell-shape in the case of many irrelevant instruments (and strong/medium endogeneity) is different from the case of a strong instrument: many irrelevant instruments yield a bell-shape around $\hat{\beta}_{OLS}$, which is far away from the true value of $\beta = 0$ (for our simulated data set) in the case of strong/medium endogeneity, whereas a strong instrument yields a bell-shape in the neighborhood of $\beta = 0$. Hoogerheide, Kaashoek and Van Dijk (2006) provide some graphical illustrations of this.

Figure 4.3 shows the marginal posterior of β in (4.13) for our nine simulated data sets; the posterior kernels are normalized over the displayed range. Notice that the graphs display fat tails in the cases of no identification, combined with a sharp peak in the case of strong endogeneity; in these cases the kernel (4.13) is approximately equal to $[(y-x\beta)'(y-x\beta)]^{-k/2}$ as $M_Z y \approx y, M_Z x \approx x$. Also note the bimodality in the case of the weak instrument and strong endogeneity; this results from the term

$$\left[\frac{(y-x\beta)'M_Z(y-x\beta)}{(y-x\beta)'(y-x\beta)}\right]^{(T-k-1)/2} = \left[1 - \frac{x'P_Z x\ (\beta-\hat{\beta}_{2SLS})^2}{y'M_x y + x'x(\beta-\hat{\beta}_{OLS})^2}\right]^{(T-k-1)/2} \tag{4.14}$$

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

$$\left[1 - \frac{r_{x,z}^2\ \left(\frac{\beta-\hat{\beta}_{2SLS}}{s_y/s_x}\right)^2}{1-r_{y,x}^2 + \left(\frac{\beta-\hat{\beta}_{OLS}}{s_y/s_x}\right)^2}\right]^{(T-2)/2} \tag{4.15}$$

with $s_y \equiv \sqrt{y'y}$ in the case of $k = 1$ instrument. In the case of one weak instrument and strong endogeneity $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{OLS}$ are in general far apart, while $r^2_{x,z}$ is small and $r^2_{y,x}$ is close to one, so that (4.15) takes very small values near $\beta = \hat{\beta}_{OLS}$ (≈ 1 for our simulated data set from the model with $\beta = 0$), whereas on both sides of $\hat{\beta}_{OLS}$ there is an interval where (4.15) is not negligible. In the cases with a strong instrument the graphs show a bell-shape; in these cases the term (4.15), converging to the very small constant $(1 - r^2_{x,z})^{(T-2)/2}$ when β becomes large (in absolute sense), makes the graph seem to be bell-shaped; also in these cases (4.15) is very small near $\beta = \hat{\beta}_{OLS}$ if $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{OLS}$ are far apart, but the large value of $r^2_{x,z}$ causes (4.15) to be only large on one relatively small interval around $\beta = \hat{\beta}_{2SLS}$, so that the graphs do not display bimodality. For a more detailed analysis comparing Bayesian and classical inference in an instrumental variable regression model we refer to Kleibergen and Zivot (2003). The highly non-normal shapes of the posterior of β in the case of weak instruments is also illustrated by Lancaster (2004).

Restricted reduced form inference

Kleibergen and van Dijk (1994b, 1998) derive the conditional posterior density of Π given β and the marginal posterior density kernel of Π. We summarize their results in two propositions:

Proposition 3: Conditional posterior of Π given β: *In the IV regression model (4.1)-(4.2) with prior (4.5) the conditional posterior density of Π given β is a Student t density with mode $\hat{\Pi} \equiv (Z'M_\varepsilon Z)^{-1}(Z'M_\varepsilon x)$, scale $s^2_{\hat{\Pi}}(Z'M_\varepsilon Z)^{-1}$ and $(T-k)$ degrees of freedom:*

$$\begin{aligned} p(\Pi|\beta, y, x, Z) &= c_2 \, |s^2_{\hat{\Pi}}(Z'M_\varepsilon Z)^{-1}|^{-1/2} \times \\ &\quad \times \left[1 + \frac{1}{T-k}(\Pi - \hat{\Pi})'(s^2_{\hat{\Pi}}(Z'M_\varepsilon Z)^{-1})^{-1}(\Pi - \hat{\Pi})\right]^{-T/2} \end{aligned} \tag{4.16}$$

where $(T-k)s^2_{\hat{\Pi}} \equiv (x - Z\hat{\Pi})'M_\varepsilon(x - Z\hat{\Pi})$ and c_2 is a scaling constant that only depends on T and k; $M_\varepsilon \equiv I - \varepsilon(\varepsilon'\varepsilon)^{-1}\varepsilon'$, with $\varepsilon \equiv y - x\beta$.

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

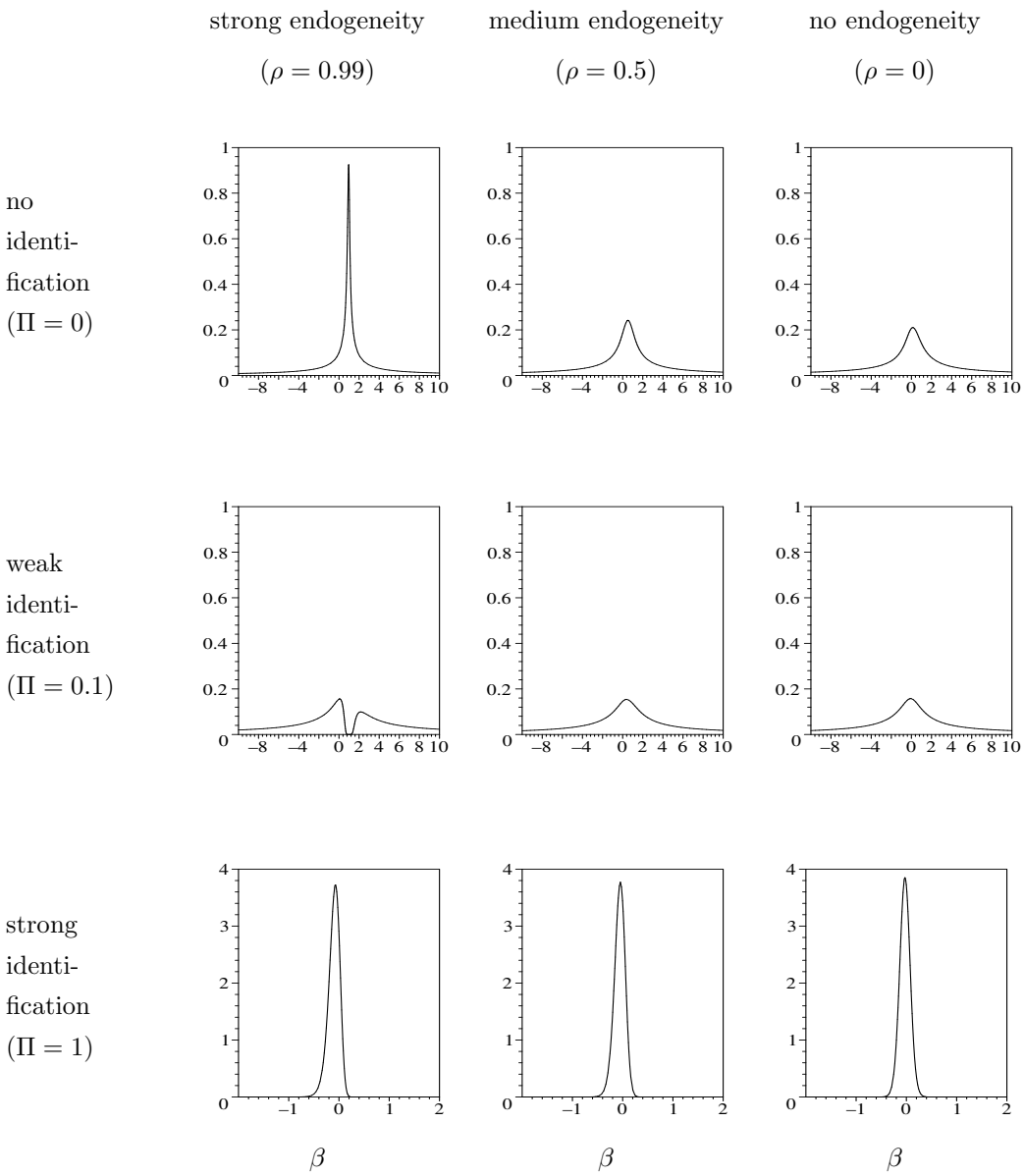

Figure 4.3: Marginal posterior kernel of β in (4.13) in IV model for nine simulated data sets; three cases of identification ($\Pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity)

Proposition 4: Marginal posterior of Π: *In the IV regression model (4.1)-(4.2) with prior (4.5) the marginal posterior density of Π is proportional to the ratio of a product of two Student t kernels in the numerator and one Student t kernel in the denominator:*

$$p(\Pi|y,x,Z) \propto \frac{[(x-Z\Pi)'(x-Z\Pi)]^{-(T-1)/2} \ (\Pi'Z'M_{[y\,x]}\,Z\Pi)^{-(T-1)/2}}{(\Pi'Z'M_xZ\Pi)^{-(T-2)/2}} \tag{4.17}$$

$$= [(x-Z\Pi)'(x-Z\Pi)]^{-(T-1)/2} \times \times(\Pi'Z'M_xZ\Pi)^{-1/2}\left(\frac{\Pi'Z'M_xZ\Pi}{\Pi'Z'M_{[y\,x]}\,Z\Pi}\right)^{(T-1)/2}, \tag{4.18}$$

known as the 2-1 poly t density.

For the overidentified case the marginal posterior distribution of Π is proper. However, for the just identified case the density kernel in (4.18) is not integrable over neighborhoods around $\Pi = 0$ (because of the term $(\Pi'Z'M_xZ\Pi)^{-1/2}$), so that this is not a proper density. Given this non-integrability, reduced form inference on Π is not possible. This result does *not* depend on the quality of the instruments *nor* on the endogeneity in the data. Only if the restriction that x is not an endogenous regressor, $\sigma_{12} = 0$, is imposed on the model *beforehand* we obtain a proper marginal density of Π. For example, specifying $p(\beta, \Pi, \sigma_{11}, \sigma_{22}) \propto \sigma_{11}^{-1/2}\sigma_{22}^{-1/2}$ and integrating out σ_{11} and σ_{22} using properties of the inverted Gamma distribution (see Zellner (1971)) yields the joint posterior of β and Π given by $p(\beta, \Pi|y,x,Z) \propto [(y-x\beta)'(y-x\beta)]^{-T/2}[(x-Z\Pi)'(x-Z\Pi)]^{-T/2}$, i.e. β and Π have independent Student t distributions with $T-1$ and $T-k$ degrees of freedom, respectively.

So, for the just identified case in the model (4.1)-(4.5) forecasting future values of x using posterior moments of Π is not possible if one uses the restricted reduced form, unless the region of integration of Π is truncated, the effect of which is not known a priori. However, it may occur that the data are such that the asymptote will not be noticed in the computations; this may happen if the mode of the joint posterior of (β, Π) occurs far away from $\Pi = 0$. Figure 4.4 shows the marginal posterior density kernel of Π in (4.18) for our nine simulated data sets. Note that each plot reveals an asymptote at $\Pi = 0$; however, for the cases of strong identification the spike near $\Pi = 0$ is very narrow and relatively far away from the bell-shaped part of the graph around $\Pi = \hat{\Pi}_{OLS}$ (≈ 0.9 for this simulated data set).

It may seem paradoxical that if equation (4.1) is excluded from the model, forecasting based on (4.2) is standard, whereas adding the extra information in equation (4.1), $y =$

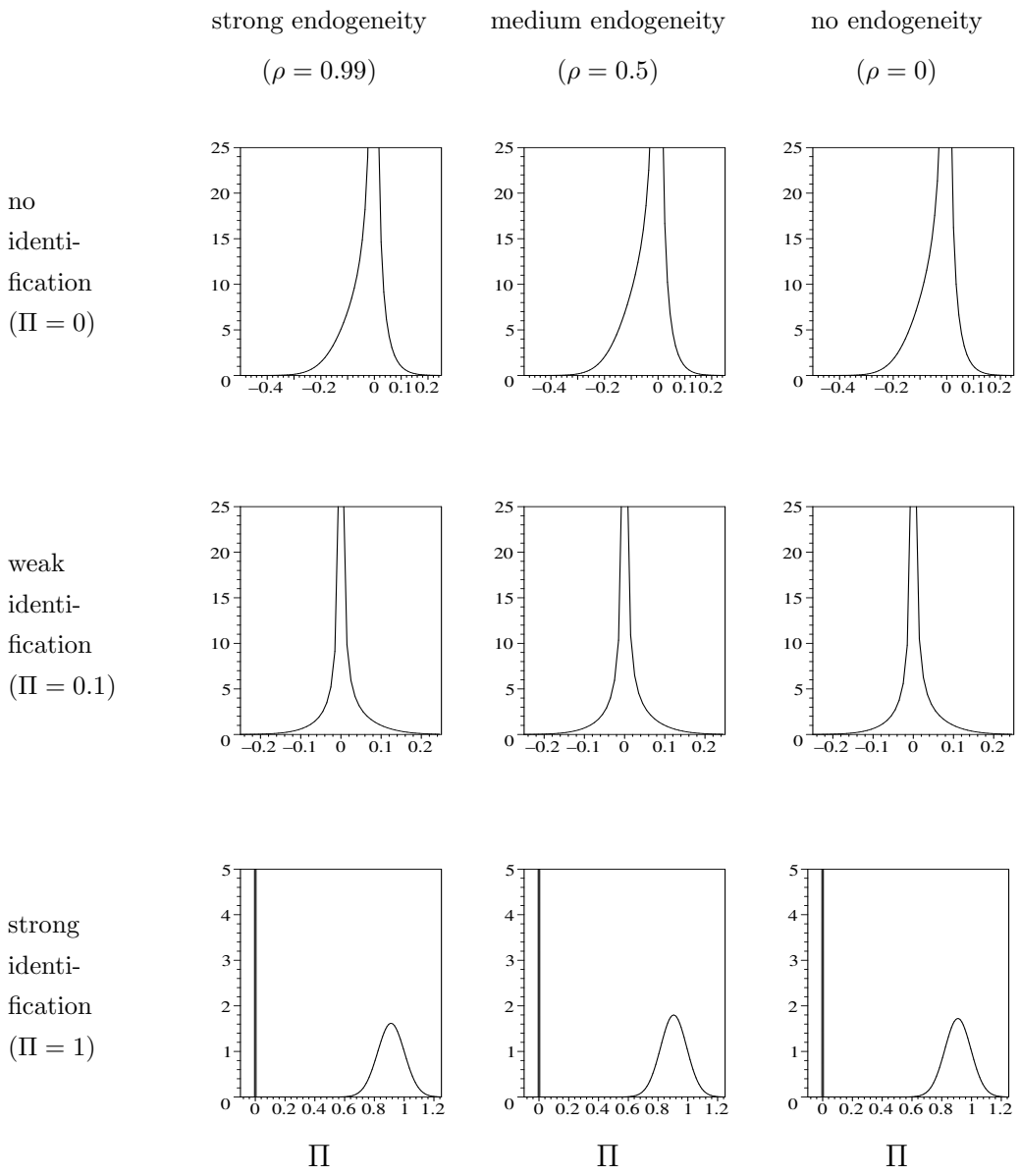

Figure 4.4: Marginal posterior kernel of Π *in (4.17) in IV model for nine simulated data sets; three cases of identification (*$\Pi = 0,\ 0.1,\ 1$ *corresponding to no, weak, strong identification) are combined with three levels of endogeneity (*$\rho = 0.99,\ 0.5,\ 0$ *corresponding to strong, medium, no endogeneity). An asymptote at* $\Pi = 0$ *occurs in each figure.*

$x\beta + \varepsilon$ with ε possibly correlated with v, makes this impossible. However, as Kleibergen and van Dijk (1994a) and Chao and Phillips (1998) point out, the flat prior for (β, Π) implies a highly informative prior for the parameters (Π_1, Π) of the restricted reduced form

$$y = Z\Pi_1 + v_1, \tag{4.19}$$

$$x = Z\Pi + v. \tag{4.20}$$

where $\Pi_1 = \Pi\beta$ and $v_1 = v\beta + \varepsilon$; in the just identified model ($k = 1$) there exists a 1-1 relationship between (β, Π) and (Π_1, Π), so that in that case it is easily derived that $p(\Pi_1, \Pi) \propto p(\beta, \Pi)\,|\partial(\beta, \Pi)/\partial(\Pi_1, \Pi)| = |\Pi|^{-1}$: the prior for (Π_1, Π) is far from non-informative for Π, as it gives infinite density to the point $\Pi = 0$.

Finally, consider the joint posterior of $\Pi = (\pi_1, \pi_2)'$ and β in (4.6) for $T = 50$ simulated data points from the model (4.1) - (4.2) with $\beta = 0$, $\sigma_{11} = \sigma_{22} = 1$, and $\rho = 0.99$ (strong endogeneity), with $k = 2$ vectors of instruments consisting of i.i.d. N(0,1) draws. Figure 4.5 shows the shape of an HPD credible set of (π_1, π_2, β) for simulated data sets with $\pi_1 = \pi_2 = 0$ (no identification), $\pi_1 = \pi_2 = 0.1$ (weak identification) and $\pi_1 = \pi_2 = 1$ (strong identification). Note that the same shapes that showed up in the 2-dimensional distributions (ridges, bimodality and nearly elliptical shapes) also occur in these 3-dimensional distributions. The bimodal posterior distribution in the case of weak identification is used in order to illustrate neural network sampling methods in chapters 2 and 3.

4.2.2 Posteriors and credible sets under Jeffreys prior

In the previous subsection it was discussed that the posterior distribution of β and Π under the flat prior has some peculiar properties. Two of these properties are that the marginal posterior of Π has an asymptote at $\Pi = 0$ (because of the term $(\Pi' Z' M_x Z \Pi)^{-1/2}$), which is non-integrable in the case of exact identification; and that the tail behavior of the marginal posterior of β depends on the number of instruments in the sense that the marginal posterior of β is improper in the case of exact identification and its tails become thinner when (possibly irrelevant) instruments are added to the model. In this subsection we consider the posterior distribution in the simple instrumental variable regression model (4.1)-(4.5) under a different prior specification, the Jeffreys prior.

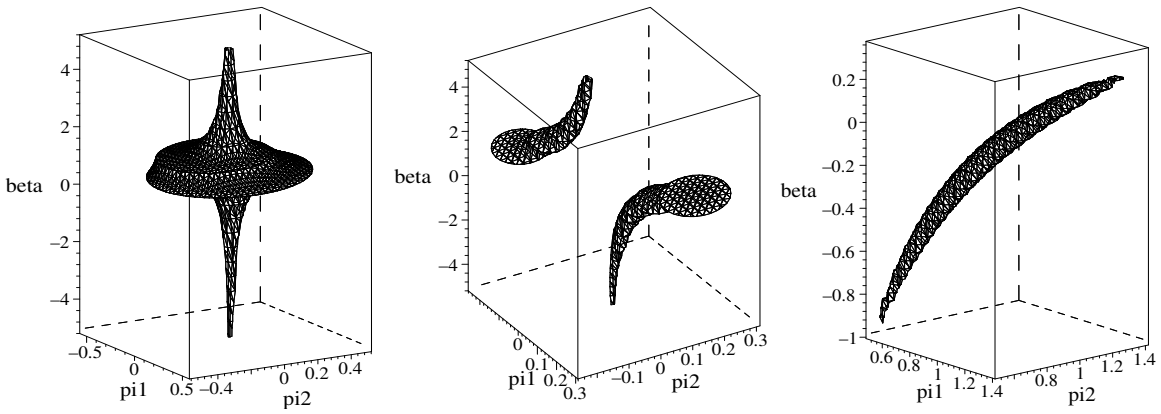

Figure 4.5: Credible sets for parameters π_1, π_2, β in IV model (4.1) - (4.2) under flat prior for simulated data sets from this model with strong endogeneity ($\rho = 0.99$) combined with either no ($\pi_1 = \pi_2 = 0$), weak ($\pi_1 = \pi_2 = 0.1$) or strong ($\pi_1 = \pi_2 = 1$) identification, respectively.

For the model (4.1)-(4.5) with one explanatory endogenous variable the Jeffreys prior, the square root of the determinant of the information matrix, is given by:

$$p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-2} (\Pi' Z' Z \Pi)^{1/2} \sigma_{22.1}^{-1/2(k-1)} \tag{4.21}$$

with $\sigma_{22.1} \equiv \sigma_{22} - \sigma_{12}^2/\sigma_{11}$ for the structural form (4.1)-(4.2) or equivalently by:

$$p(\beta, \Pi, \Omega) \propto |\Omega|^{-2} (\Pi' Z' Z \Pi)^{1/2} ((\beta : 1)\Omega^{-1}(\beta : 1)')^{1/2(k-1)} \tag{4.22}$$

for the corresponding restricted reduced form (4.3)-(4.4); see for example Appendix A of Hoogerheide, Kleibergen and Van Dijk (2006) 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 term $(\Pi' Z' Z \Pi)^{1/2}$ 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 $((\beta : 1)\Omega^{-1}(\beta : 1)')^{1/2(k-1)}$ factor in the prior influences the tail behavior of the marginal posterior of β and makes it independent of the number of instruments such that it has Cauchy type tails.

Note that for $k = 1$ instrument the Jeffreys prior (4.21) reduces to

$$p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-2}|\Pi|, \tag{4.23}$$

which is simply the flat prior of Drèze (1976) in (4.5) with $h = 4$ multiplied with $|\Pi|$. A strange interpretation of this Jeffreys prior would be to say that one a priori expects Π to be large (in absolute sense). An intuitively more 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 peculiar properties resulting from the flat 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 $((\beta : 1)\Omega^{-1}(\beta : 1)')^{1/2(k-1)}$ 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 product $\Pi\beta$ that causes points (Π, β) with Π and β both attaining (extremely) large values to have small posterior probability.

Figure 4.6 shows the joint posterior of (Π, β) under the Jeffreys prior for the same nine simulated data sets that are used in the previous subsection in order to illustrate the shapes of posteriors under the flat prior. This is the case with $k = 1$ instrument, where the posterior of (Π, β) under the Jeffreys is given by the posterior under the flat prior multiplied by the absolute value of Π (and with exponent $-T/2$ changed into $-(T+1)/2$). In the case of a strong instrument there seems to be little difference between the posterior under the Jeffreys or flat prior. Obviously, the $|\Pi|$ factor implies that there is less posterior probability mass in neighborhoods of $\Pi = 0$ than under the flat prior, which clearly affects the posterior in the six cases of weak or no identification. In these cases the contour plots show bimodality with posterior probability mass on both sides of the line $\Pi = 0$. The factor $|\Pi|$ implies that the marginal posterior of Π has no asymptote at $\Pi = 0$. However, it should be noted that for example in the case of a weak instrument and no endogeneity the *marginal* posterior of Π does not drop in neighborhoods of $\Pi = 0$ either: for $\Pi \to 0$ the lower values of the posterior density kernel $p(\Pi, \beta|y, x, Z)$ are compensated by the fact that for $\Pi \to 0$ the posterior $p(\Pi, \beta|y, x, Z)$ 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. This phenomenon is reflected by the contour plot in the second row and third column of Figure 4.6.

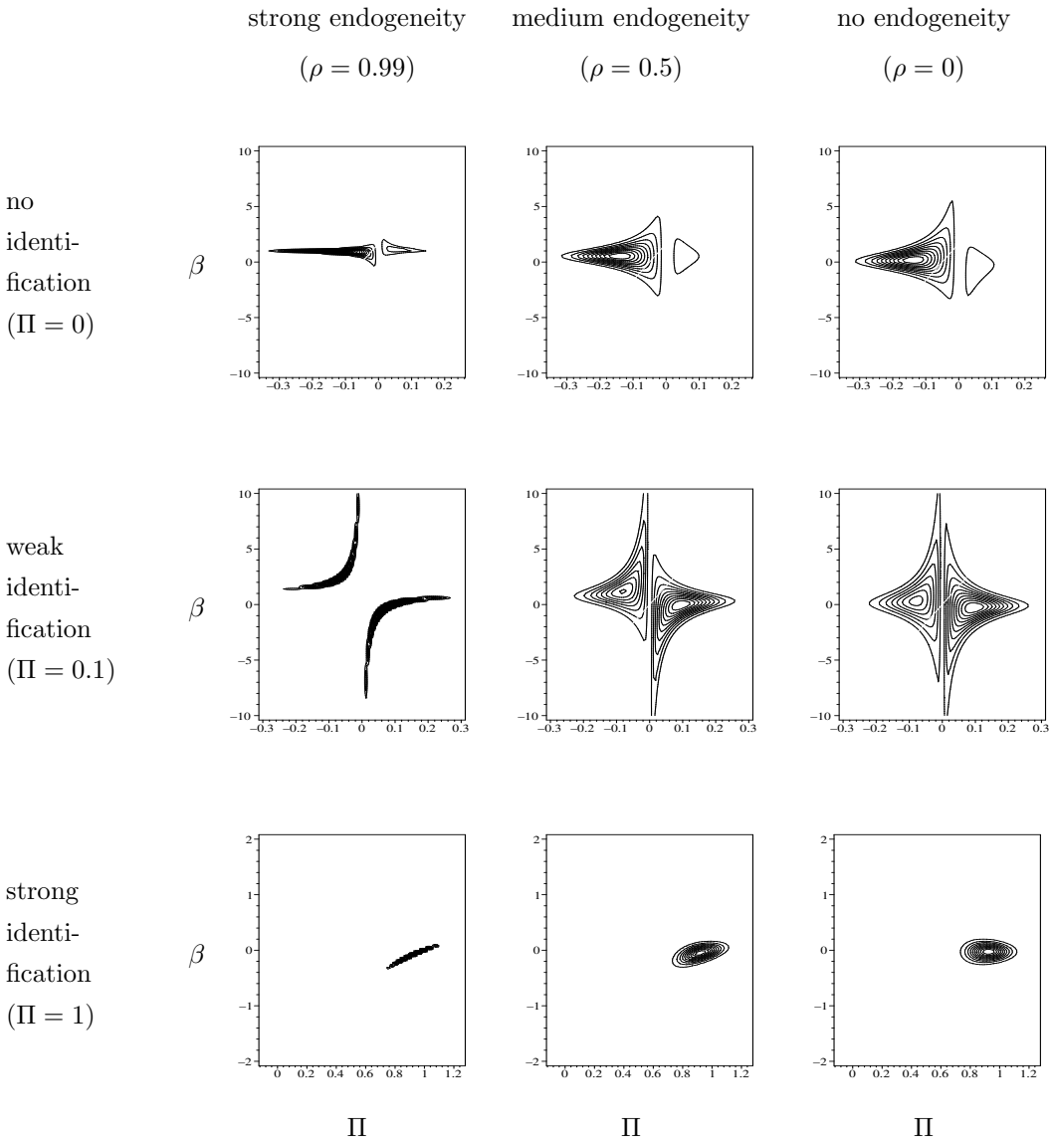

Figure 4.6: Contour plots in the $\Pi \times \beta$ plane: joint posterior kernel of Π and β in IV model under Jeffreys prior for nine simulated data sets; three cases of identification ($\Pi = 0,\ 0.1,\ 1$ corresponding to no, weak, strong identification) are combined with three levels of endogeneity ($\rho = 0.99,\ 0.5,\ 0$ corresponding to strong, medium, no endogeneity)

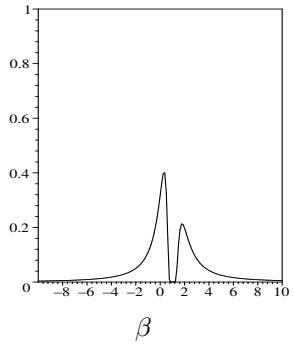

Figure 4.7: Marginal posterior of β under Jeffreys prior for simulated data set with weak identification ($\Pi = 0.1$) and strong endogeneity ($\rho = 0.99$)

Figure 4.7 shows the marginal posterior of β under the Jeffreys prior in the case of a weak instrument and strong endogeneity: this shows that the tails are thinner than under the flat prior, reflecting that this is a proper posterior, while the posterior under the flat prior is not. The bimodality shows that the Jeffreys prior may still result in a bimodal marginal posterior of β.

The Jeffreys prior 'cures' some of the peculiar properties of posterior distributions under the flat prior, the asymptote of the marginal posterior of Π at $\Pi = 0$ and the dependence of the tail behavior of the marginal posterior of β on the number of (possibly irrelevant) instruments. However, posterior distributions under the Jeffreys prior may still show non-elliptical shapes such as bimodality.

For the case of one explanatory endogenous variable an explicit formula exists for the marginal posterior of β, see *e.g.* Kleibergen and Zivot (2003). If there are m explanatory endogenous variables with $m > 1$, then sampling methods are required. Kleibergen and van Dijk (1998) and Kleibergen and Paap (2002) have derived specific importance sampling and Metropolis-Hastings algorithms for models in which a reduced rank restriction is imposed on a parameter matrix, for example the instrumental variables regression model and the vector error correction model (VECM). However, these may require the evaluation of a determinant of a $(m+1)k \times (m+1)k$ Jacobian matrix, which may be numerically cumbersome, especially in the case of many instruments. If these sampling methods are not applicable in certain cases, the possibility of (highly) non-elliptical shapes of the posterior distributions under the Jeffreys prior implies that neural network sampling methods may be useful tools in a Bayesian analysis of an IV model under the Jeffreys prior, possibly after some parameter transformations. This is left as a topic for further research.

4.3 Hierarchical prior of Chamberlain and Imbens (1996)

In the previous section some peculiar properties are considered of posterior distributions under the flat prior, and an intuitive explanation is given of how the Jeffreys prior 'remedies' some of these properties. In this sense the Jeffreys prior is a 'regularization prior' that 'cures' strange properties occurring under the flat prior. In this section we briefly discuss the hierarchical prior of Chamberlain and Imbens (1996), which can also be considered as a 'regularization prior'. This prior specification is inspired by the drawback of inference under the flat prior that many irrelevant instruments result in a tight marginal posterior of β, even though no information on β is present in the data. Kleibergen and Zivot (2003) also consider the behavior of the marginal posterior of β under the flat and Jeffreys prior in the case of many irrelevant instruments, and show that the posterior under the Jeffreys prior is (relatively) insensitive to the addition of many irrelevant instruments. First, we summarize the prior specification suggested by Chamberlain and Imbens (1996); then we compare it to the Jeffreys prior.

Chamberlain and Imbens (1996) argue as follows. Consider the IV model with structural form (4.1)-(4.2) and restricted reduced form (4.3)-(4.4). When k, the number of instruments in Z, is large, a choice for a flat prior distribution is in fact very informative, because the prior distribution dogmatically asserts that the instrumental variables Z are *collectively* very powerful predictors of the explanatory endogenous variable x. In this case it is therefore important to restrict the variability of the parameters Π. Therefore a structure is imposed on the prior distribution in the form of a hierarchical (nested) model. It is assumed that the elements of the vector Π, π_j $(j = 1, \ldots, k)$, obey the distribution:

$$\pi_j | \alpha, \sigma_\pi^2 \overset{i.i.d.}{\sim} N(\alpha, \sigma_\pi^2) \qquad (j = 1, \ldots, k) \tag{4.24}$$

where α and σ_π^2 are hyperparameters. For β, Ω and the hyperparameter α improper priors are specified:[5]

$$p(\Pi, \beta, \Omega, \alpha, \sigma_\pi^2) \propto p(\Pi | \alpha, \sigma_\pi^2) p(\sigma_\pi^2) |\Omega|^{-(k+3)/2}, \tag{4.25}$$

[5]The model we describe is a slight modification of the model used by Chamberlain and Imbens (1996). Chamberlain and Imbens (1996) consider a model containing also exogenous variables that are included in all equations; the parameters of these exogenous variables are included in the π_j, such that π_j $(j = 1, \ldots, k)$ and α are vectors and σ_π^2 is replaced by a covariance matrix. This is possible and plausible in their particular example, since and these included and excluded exogenous variables correspond to the same 'units'.

where the prior density $p(\sigma_\pi^2)$ is given by an inverted Wishart density:

$$(\sigma_\pi^2)^{-1} \sim Wishart(h, H),$$

which is an inverted Gamma distribution in this case of a scalar. A Gibbs sampler can be used to sample from the posterior resulting from this hierarchical prior, see the appendix of Chamberlain and Imbens (1996). A conventional diffuse, but improper, prior for σ_π^2 would correspond to $h = 0, H^{-1} = 0$. However, Chamberlain and Imbens (1996) show that this results in an improper posterior in which the event $\sigma_\pi^2 = 0$ has probability one. Therefore a proper prior is required. The value of h is chosen as the smallest value such that there is probability one that σ_π^2 is nonsingular, that is $h = 1$. The value of H^{-1} is chosen as $H^{-1} = C \cdot h \cdot \text{cov}(\hat{\pi})$, where $\text{cov}(\hat{\pi}) = \sum_{j=1}^{k}(\hat{\pi}_j - \bar{\hat{\pi}})'(\hat{\pi}_j - \bar{\hat{\pi}})/k$, $\bar{\hat{\pi}} = \sum_{j=1}^{k} \hat{\pi}_j/k$ with $\hat{\pi}_j$ $(j = 1, \ldots, k)$ OLS estimates in the first stage regression. (The use of the data dependent $\text{cov}(\hat{\pi})$ is not essential; it is merely a convenient normalization.)

The choice of the positive constant C is crucial for obtaining useful results from this hierarchical model: if C is chosen very large, this corresponds essentially to a flat prior on Π, and results will be similar to the Drèze approach, yielding misleadingly small HPD regions in the case of many superfluous instruments. On the other hand, if C is too small, the posterior distribution of β will have misleadingly large variance when the 'true' π_j $(j = 1, \ldots, k)$ have non-zero values on both sides of 0, but are 0 on average. For in that case, the instruments contain information, but restricting the π_j $(j = 1, \ldots, k)$ to be all approximately equal – by imposing a prior on σ_π^2 that strongly favors very small σ_π^2 – prevents the method to use the information in the instruments. (Note that if the 'true' values of the π_j $(j = 1, \ldots, k)$ would have a non-zero average, a non-zero value of α could cause the posterior of Π to be located away from 0, thus yielding a tight posterior of β despite the (unwisely chosen) prior on σ_π^2.)

The constant C should therefore be chosen in such a way that it puts prior mass both close to zero (to guard against misleading inference in the case of many superfluous instruments) and on large values of σ_π^2 (to guard against loss of information contained in the instruments). In specific applications, the choice of C may be based on the relative concern with both cases. A careful 'tuning' of the choice of C (using a simulated data set containing relevant instruments and a simulated data set containing only irrelevant instruments) is therefore necessary.

It is immediately clear that the 'tuning' of the choice of C in the hierarchically based prior in the approach of Chamberlain and Imbens (1996) is a major disadvantage as

compared to the approach under the Jeffreys prior, which does not require such 'tuning'. Further, under the Jeffreys prior one does not make the assumption of normally distributed elements of Π; it is not immediately clear that this assumption does not have undesired effects on the results.

It should be noted that the approach of Chamberlain and Imbens (1996) has the advantage that the hierarchical prior is not necessarily data dependent, while the Jeffreys prior generally is (because of $Z'Z$ occurring in it); and that a straightforward Gibbs sampler can be used to sample from its posterior. However, the two disadvantages mentioned above clearly seem to outweigh these advantages.

4.4 Concluding remarks

In this chapter we have considered how shapes of posteriors in the IV model under the flat or Jeffreys prior depend on the level of endogeneity and instrument strength. Further, it is considered how the Jeffreys prior 'remedies' two of the peculiar properties of the posterior under the flat prior, the asymptote of the marginal posterior of Π at $\Pi = 0$ and the dependence of the tail behavior of the marginal posterior of β on the number of instruments in the sense that the marginal posterior of β is improper in the case of exact identification, whereas its tails become thinner when (possibly irrelevant) instruments are added to the model.

For the case of one explanatory endogenous variable an explicit formula exists for the marginal posterior of β under the Jeffreys prior, see *e.g.* Kleibergen and Zivot (2003). If there are m explanatory endogenous variables with $m > 1$, then sampling methods are required. Kleibergen and van Dijk (1998) and Kleibergen and Paap (2002) have derived importance sampling and Metropolis-Hastings algorithms that are specifically designed for reduced rank regression models such as the IV regression model. However, in the case of many instruments these may require the evaluation of a determinant of a huge Jacobian matrix, which may be numerically cumbersome. If these sampling methods are not applicable in certain cases, the possibility of (highly) non-elliptical shapes of the posterior distributions under the Jeffreys prior implies that neural network sampling methods may be useful tools in a Bayesian analysis of an IV model under the Jeffreys prior, possibly after some parameter transformations. This is left as a topic for further research.

Finally, the hierarchical prior of Chamberlain and Imbens (1996) is briefly discussed, which can also be considered as a 'regularization prior' that 'cures' strange posterior

properties occurring under the flat prior. Unlike the approach under the flat prior, the approach of Chamberlain and Imbens (1996) is also capable of only resulting in tight HPD regions for β in the case of data sets that contain information on β, just like the approach using the Jeffreys prior. The hierarchically based prior in the approach of Chamberlain and Imbens (1996) requires the 'tuning' of some prior variance (or covariance matrix), which is obviously a major disadvantage as compared to the approach under the Jeffreys prior. It should be noted that the approach of Chamberlain and Imbens (1996) has the advantage that the hierarchical prior is not necessarily data dependent, while the Jeffreys prior generally is, and that a straightforward Gibbs sampler can be used to sample from the corresponding posterior. However, the disadvantage of the 'tuning' of a prior variance, and the sensitivity of posterior results to the choice of this prior variance, clearly suggests that the use of the Jeffreys prior is preferable in most situations.

Chapter 5

An instrumental variables regression model for return on education: Angrist-Krueger reconsidered

Chapter 5 is based on Hoogerheide, Kleibergen and Van Dijk (2006) and Hoogerheide and Van Dijk (2006b).

5.1 Introduction

In this chapter we consider an instrumental variables regression model due to Angrist and Krueger (1991) for the effect of education on income. Because of the endogeneity of the years of education and income, Angrist and Krueger use instruments that are obtained from the quarter of birth. It is hard to find instruments that are correlated with education but uncorrelated with unobserved 'ability' which explains both the education and income. Estimating the return on education is therefore a non-trivial matter. The instruments that are based on the quarter of birth exploit that students born in different quarters have different average education. This results since most school districts require students to have turned age six by January 1 of the year they enter school and 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.[1]

For quarter of birth to be a valid instrument it should only influence income through its effect on education. This is a plausible assumption, as one's birthday is unlikely to be correlated with personal attributes other than age at school entry.[2] Moreover, Angrist and Krueger (1991) do not find evidence of an effect of quarter of birth on the years of education for college graduates. Compulsory schooling laws do not compel persons to attend school beyond high school, so if such evidence were found it would mean that there were also different reasons (like characteristics of one's family or personal attributes such as intelligence or 'ability' in general) causing an effect of quarter of birth on education, and probably also a direct effect of quarter of birth on income. The fact that no such evidence was found strengthens the idea that quarter of birth only influences education through the compulsory school attendance, and has no direct influence on income.

The strength of these instruments clearly depends on the fraction of students that immediately leave school when it is permitted. This is, however, only a small part of the total population of students since most students do not immediately leave school when it is allowed and some leave school before they attain the legal dropout age. Angrist and Krueger (1991) mention several factors that influence the size of the latter group. Compulsory schooling laws allow certain officers to take children into custody and/or punish a child's parents if a child does not attend school; and child labor laws restrict or prohibit children of compulsory school age from participating in the work force, the main alternative to attending school. There are, however, exemptions to compulsory schooling laws: students are exempt from compulsory school attendance if they have a high school degree; and in many states there are exemptions for children suffering from physical or mental disabilities, or if they live far from school.

Alongside that the quarter of birth only affects the years of education for a small fraction of the student population, its influence is also limited since it only implies a

[1]The assumption that the birthday cutoff is January 1 is not crucial; the main point is that children with different birthdays are allowed to leave school after different amounts of education. Angrist and Krueger (1992) mention the different cutoffs for several states in 1955, which influenced a different group of children than our data set. If these cutoffs would be the same in the period 1936-1945 as in 1955, this would still not be alarming. The only influence of the exact birthday cutoff is that instruments based on quarter of birth are obviously somewhat more powerful when the birthday cutoff is at the beginning/end of a quarter.

[2]Bound, Jaeger and Baker (1995) criticized this assumption; the criticism of Bound, Jaeger and Baker (1995) will be discussed in section 5.5.

maximum difference of one year over the different quarters which is small compared to the overall variation in the education spell. Quarter of birth is therefore expected to be a weak instrument. Bound, Jaeger and Baker (1995), for example, show that randomly generated instruments, designed to match the data of Angrist and Krueger (1991), yield results remarkably similar to those based on the actual instruments. Staiger and Stock (1997) also show that inference on the return on education is strongly affected by the weakness of the quarter of birth instruments. Hence, although the quarter of birth seems a plausible source for constructing instruments, we should be careful with interpreting the results because of the weakness of the instruments.

Section 5.2 describes the particular model and data that we use. In section 5.3 and 5.4 classical and Bayesian results are given, respectively. In sections 5.2 - 5.4 it is assumed that the assumptions made Angrist and Krueger (1991) are satisfied by the data. In section 5.5 some assumptions made by Angrist and Krueger (1991) are investigated. Section 5.6 gives conclusions.

5.2 Model and data

Angrist and Krueger (1991) use data sets concerning men born in the USA in the years 1920-1929, 1930-1939 or 1940-1949, and consider several model specifications. We use a subset of the data used by Angrist and Krueger (1991): a data set on income, years of education and state/quarter/year of birth consisting of 329,509 men born in the USA in the years 1930-1939.[3] We use the following model:

$$\tilde{y}_i = \tilde{x}_i\beta + \sum_{j=1}^{9} D_{j,i}^{y}\,\delta_j^{y} + \sum_{t=1}^{S-1} D_{t,i}^{s}\,\delta_t^{s} + \pi_1 + \tilde{\varepsilon}_i \qquad i = 1,\ldots,T \tag{5.1}$$

$$\begin{aligned}\tilde{x}_i = \;& \sum_{j=1}^{9} D_{j,i}^{y}\gamma_j^{y} + \sum_{t=1}^{S-1} D_{t,i}^{s}\gamma_t^{s} + \pi_2 \\ & + \sum_{t=1}^{S}\sum_{h=2}^{4} D_{t,i}^{s}D_{h,i}^{q}\,\pi_{th}^{sq} + \sum_{j=1}^{9}\sum_{h=2}^{4} D_{j,i}^{y}D_{h,i}^{q}\,\pi_{jh}^{yq} + \tilde{v}_i\end{aligned} \tag{5.2}$$

where $\tilde{y}_i$ is the logarithm of the weekly wage of person i in 1979, $\tilde{x}_i$ is the number of *completed* years of education by person i, and the parameter of interest is the return on education β. The dummy variables $D_{t,i}^{s}$, $D_{j,i}^{y}$, $D_{h,i}^{q}$ are equal to 1 if individual i was born in state t, year $1929+j$, quarter h, and equal to 0 otherwise, respectively. S is the number of states of birth, *i.e.* $S = 51$ (including the District of Columbia) if we use all states. We however also consider four subsamples for which we divide the US into four regions that are also used by the US Census Bureau, the source of the data. The states and

[3]The source of the data is the 1980 Census, 5 percent Public Use Sample.

Table 5.1: US Census Bureau Regions

	Census region	number of observations	number of states	states (including D.C.)
1.	Northeast	84484	9	Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont.
2.	Midwest	102267	12	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin.
3.	South	114391	17	Alabama, Arkansas, Delaware, D.C., Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia.
4.	West	28367	13	Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming.
	USA	329509	51	

numbers of observations in each region are given by Table 5.1. The coefficients π_1 and π_2 are the constant terms; $\tilde{\varepsilon}_i$ and $\tilde{v}_i$ are disturbances that are assumed to be jointly normal distributed and independent across individuals.

The state and year dummies $D^s_{t,i}$ and $D^y_{j,i}$ are included in both equations since state and year of birth both influence the education spell and income. The year dummies in the wage equation (5.1) incorporate the effect of age (measured in years) on income.

The exogenous variables that are excluded from the wage equation (5.1) are the interactions of state and quarter of birth dummies, and interactions of year and quarter of birth dummies. The interacted state and quarter of birth dummies reflect that the influence of the quarter of birth on education may differ between states which results since compulsory education laws differ between states. The legal dropout age varies between 16, 17 and 18 years and in some states students have to finish the school term. The rules concerning exemptions from the compulsory school attendance vary as well across states. The average number of years of education that students desire also differs between states (see Tables 5.2 and 5.3, which show that on average men born in the Southern region in the period 1930-1939 have less education than men born in the other regions); the more years of education that students on average want to attend, the smaller the fraction of

students that leave school as soon as the law allows it, and the smaller the coefficients at the interacted state and quarter of birth dummies. Note that π_{th}^{sq} is interpreted as the effect of the h-th ($h = 2, 3, 4$) quarter on education in state s in 1939, i.e. the difference in the years of education between men born in the h-th quarter and the first quarter in 1939 (on average).

The interacted year and quarter of birth dummies reflect that the influence of the quarter of birth on education may change over time. For example, the average number of years of education that students desire may change over time. In fact the average number of years of education has increased from 1930 to 1939, see Table 5.4.[4] Note that π_{jh}^{yq} ($j = 1, \ldots, 9; h = 2, 3, 4$) is interpreted as the difference in the effect of the h-th ($h = 2, 3, 4$) quarter on education between the year $1929 + j$ and 1939, i.e. the difference between the differences in years of education between men born in the h-th quarter and the first quarter between the year $1929 + j$ and 1939 (on average).

Model (5.1)-(5.2) reads in matrix notation:

$$\tilde{y} = W\Pi_1 + \tilde{X}\beta + \tilde{\varepsilon} \tag{5.3}$$

$$\tilde{X} = W\Pi_2 + \tilde{Z}\Pi + \tilde{V} \tag{5.4}$$

where $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_T)'$, $\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_T)'$, $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_T)'$, $\tilde{V} = (\tilde{v}_1, \ldots, \tilde{v}_T)'$; W is the $T \times (S + 9)$ matrix of year and state of birth dummies and a constant term with rows $W_i = (D_{1i}^y, \ldots, D_{9i}^y, D_{1i}^s, \ldots, D_{S-1,i}^s, 1)$, $\tilde{Z}$ is the $T \times 3(S+9)$ matrix with rows Z_i containing the state-and-quarter of birth and year-and-quarter of birth interactions $D_{ti}^s D_{hi}^q$, $D_{ji}^y D_{hi}^q$ ($t = 1, \ldots, S$; $h = 2, 3, 4$; $j = 1, \ldots, 9$). The parameter vectors are the $(S + 9) \times 1$ vectors $\Pi_1 = (\delta_1^y, \ldots, \delta_9^y, \delta_1^s, \ldots, \delta_{S-1}^s, \pi_1)'$, $\Pi_2 = (\gamma_1^y, \ldots, \gamma_9^y, \gamma_1^s, \ldots, \gamma_{S-1}^s, \pi_2)'$ and the $3(S+9) \times 1$ vector Π containing the coefficients π_{th}^{sq}, π_{jh}^{yq} ($t = 1, \ldots, S$; $h = 2, 3, 4$; $j = 1, \ldots, 9$).

We respecify (5.3)-(5.4) as:

$$y = X\beta + \varepsilon \tag{5.5}$$

$$X = Z\Pi + V \tag{5.6}$$

where y, X, Z (and the error terms ε, V) contain the residuals of $\tilde{y}$, $\tilde{X}$, $\tilde{Z}$ (and $\tilde{\varepsilon}$, $\tilde{V}$) after regression on W; that is, the observations are 'corrected' for differences in mean

[4] Angrist and Krueger (1991) conclude that as average income in 1979 is approximately equal across birth years 1930-1939, age has no or little influence on income for men between 40 and 49 years old. However, as the average education has increased over years of birth 1930-1939, age may very well have a positive effect that is (on average) compensated by the lower level of education. Note that this does not immediately imply that the variable age should be included in the model, as the year dummies already incorporate the effect of age (measured in years).

Table 5.2: Summary statistics of education and wage per region

	Census region	number of observations	education* average	st.dev.	% ≤ 9	% ≤ 10	log weekly wage average	st.dev.
1.	Northeast	84484	13.27	3.12	9.4%	14.2%	5.96	0.65
2.	Midwest	102267	13.06	2.99	10.0%	14.6%	5.97	0.66
3.	South	114391	11.93	3.52	22.0%	28.0%	5.77	0.71
4.	West	28367	13.63	3.01	6.5%	10.1%	6.00	0.65
	USA	329509	12.77	3.28	13.7%	18.7%	5.90	0.68

*In most states people born in 1930-1939 were obliged to enter school at age 5 or 6 and allowed to leave school at age 16 having completed 9 or 10 years of education.

across years and states.[5] The restricted reduced form corresponding to the structural form (5.5)-(5.6) is given by:

$$y = Z\Pi\beta + u \tag{5.7}$$

$$X = Z\Pi + V \tag{5.8}$$

where $u \equiv V\beta + \varepsilon$. The error terms in the structural form and the restricted reduced form have covariance matrix Σ and Ω, i.e. $(\varepsilon_i, v_i)' \sim N(0, \Sigma)$ and $(u_i, v_i)' \sim N(0, \Omega)$, with

$$\Sigma = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

$$\Omega = \begin{pmatrix} \omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} = \begin{pmatrix} 1 & \beta' \\ 0 & I \end{pmatrix} \Sigma \begin{pmatrix} 1 & 0 \\ \beta & I \end{pmatrix} = \begin{pmatrix} \sigma_{11} + 2\Sigma_{21}\beta + \beta'\Sigma_{22}\beta & \beta'\Sigma_{22} + \Sigma_{12} \\ \Sigma_{22}\beta + \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Section 5.3 shows results for two classical methods, two-stage least squares (2SLS) and limited information maximum likelihood (LIML). In section 5.4 the results are given for Bayesian methods, using either a flat or Jeffreys prior.

[5]In classical inference the Frisch-Waugh-Lovell theorem implies the equivalence of the results from (5.3)-(5.4) and (5.5)-(5.6). In Bayesian inference this results since specifying a flat prior for Π_1 and Π_2 and integrating out Π_1 and Π_2 in the model (5.3)-(5.4) yields the same posterior as considering the model (5.5)-(5.6) (upto some factor that is neglectable if the number of observations T is much larger than $S+9$, the dimension of W_i, as is the case throughout this chapter).

Table 5.3: Summary statistics of education and wage per state of birth

	number of	education				log weekly wage	
state	observations	average	st.dev.	% ≤ 9	% ≤ 10	average	st.dev.
Alabama	8536	11.71	3.46	23.1	29.5	5.72	0.76
Alaska	78	13.47	3.12	7.7	15.4	6.09	0.83
Arizona	1066	13.11	3.27	11.5	16.0	5.96	0.62
Arkansas	5794	11.85	3.41	21.4	27.8	5.77	0.70
California	11078	13.87	2.93	4.6	7.8	6.04	0.65
Colorado	2818	13.32	3.11	9.1	12.8	5.95	0.65
Connecticut	3844	13.31	3.08	9.7	14.3	5.96	0.63
Delaware	598	12.30	2.94	15.7	21.2	5.85	0.61
D.C.	1237	13.83	3.12	6.4	10.6	6.01	0.65
Florida	3913	12.68	3.35	14.4	19.9	5.78	0.69
Georgia	8411	11.50	3.51	24.8	31.4	5.68	0.72
Hawaii	246	13.23	3.06	10.2	14.2	5.97	0.69
Idaho	1599	13.54	3.02	7.5	11.1	5.95	0.64
Illinois	18375	13.35	3.00	8.0	12.6	6.03	0.65
Indiana	8918	12.77	2.87	10.8	16.0	5.94	0.63
Iowa	6699	13.14	2.96	9.3	12.5	5.91	0.70
Kansas	4807	13.44	2.96	7.7	10.5	5.91	0.67
Kentucky	8933	11.27	3.55	30.7	36.8	5.80	0.70
Louisiana	5975	12.07	3.61	20.0	25.6	5.84	0.71
Maine	2424	12.35	3.10	17.0	22.0	5.75	0.64
Maryland	4139	12.44	3.27	16.7	23.4	5.88	0.67
Massachusetts	9955	13.47	3.16	8.9	13.1	5.95	0.64
Michigan	14077	13.00	2.89	9.4	14.9	6.03	0.62
Minnesota	7170	13.19	3.03	10.0	13.5	5.97	0.67
Mississippi	5864	11.49	3.73	25.9	32.6	5.68	0.75
Missouri	9274	12.69	3.18	14.8	19.6	5.90	0.69
Montana	1407	13.38	3.01	8.0	12.2	5.91	0.70
Nebraska	3488	13.34	2.96	7.5	11.1	5.92	0.70
Nevada	308	13.48	2.95	8.1	11.0	5.99	0.73
New Hampshire	1200	12.59	3.15	16.6	21.0	5.80	0.64
New Jersey	8964	13.43	3.11	8.3	13.1	6.00	0.66
New Mexico	1325	12.59	3.41	15.1	19.9	5.85	0.61
New York	29015	13.70	3.16	7.1	11.4	6.01	0.66
North Carolina	10798	11.70	3.43	24.0	30.5	5.66	0.71
North Dakota	2028	12.94	3.28	15.4	19.5	5.93	0.70

Table 5.3 (continued)

	number of	education				log weekly wage	
state	observations	average	st.dev.	% ≤ 9	% ≤ 10	average	st.dev.
Ohio	17070	12.95	2.95	10.1	15.4	5.97	0.63
Oklahoma	6950	13.00	3.11	11.4	15.9	5.90	0.66
Oregon	2127	13.65	2.86	5.2	8.9	5.99	0.62
Pennsylvania	26385	12.84	2.97	10.8	16.3	5.93	0.62
Rhode Island	1698	12.91	3.21	15.1	20.8	5.85	0.67
South Carolina	5245	11.30	3.58	27.3	34.6	5.61	0.75
South Dakota	1754	13.18	3.23	12.8	15.3	5.90	0.67
Tennessee	8335	11.54	3.51	27.0	32.8	5.75	0.71
Texas	15932	12.67	3.62	15.5	20.1	5.87	0.70
Utah	1999	13.94	3.04	5.5	9.7	6.01	0.61
Vermont	999	12.40	3.18	18.2	22.4	5.73	0.67
Virginia	7319	11.47	3.62	27.3	34.0	5.73	0.71
Washington	3610	13.66	2.90	5.4	8.7	6.04	0.64
West Virginia	6412	11.81	3.16	22.5	28.9	5.85	0.64
Wisconsin	8607	12.96	2.94	10.2	14.5	5.93	0.66
Wyoming	706	13.60	3.02	6.9	10.9	5.99	0.68

Table 5.4: Summary statistics of education and wage per year of birth

	number of	education				log weekly wage	
year	observations	average	st.dev.	% ≤ 9	% ≤ 10	average	st.dev.
1930	33602	12.46	3.44	17.2	22.6	5.90	0.69
1931	30583	12.59	3.38	15.8	20.8	5.91	0.69
1932	32211	12.63	3.40	15.9	21.1	5.90	0.69
1933	30751	12.69	3.35	14.9	20.1	5.90	0.68
1934	31916	12.72	3.32	14.4	19.7	5.90	0.69
1935	32773	12.78	3.26	13.7	18.6	5.89	0.69
1936	32676	12.84	3.20	12.6	17.6	5.90	0.67
1937	33969	12.90	3.17	11.7	16.5	5.90	0.66
1938	35223	12.99	3.15	11.1	15.8	5.90	0.67
1939	35805	13.03	3.13	10.6	15.5	5.90	0.66

5.3 Classical approaches

In this section we first briefly summarize two well-known classical single equation estimators for β, two-stage least squares (2SLS) and limited information maximum likelihood (LIML); for extensive discussions of classical single equation procedures the reader is referred to Hausman (1983) or Phillips (1983). After that the results are discussed of applying the 2SLS and LIML methods to the Angrist-Krueger IV model for data of the US and the four Census regions.

In the two-stage least squares method, due to Theil (1953) and Basmann (1957), an estimator of β is obtained by the following two steps. First, an estimate of Π in (5.6) is obtained by OLS: $\hat{\Pi}_{OLS} = (Z'Z)^{-1}Z'X$. Second, an estimate of β is obtained by OLS of y on $Z\Pi$ in (5.7) with Π replaced by $\hat{\Pi}_{OLS}$: $\hat{\beta}_{2SLS} = (X'Z\,(Z'Z)^{-1}Z'X)^{-1}X'Z\,(Z'Z)^{-1}Z'y$. The 2SLS estimator $\hat{\beta}_{2SLS}$ is a consistent estimator of β that is asymptotically normal distributed with covariance matrix $(1/T)\sigma_{11}(\Pi'\Sigma_Z\Pi)^{-1}$, where $\Sigma_Z = \text{plim}_{T\to\infty}(1/T)Z'Z$, under the conditions that β is identified and instruments are not too weak. Staiger and Stock (1997) explore a case of weak instruments, defined as $\Pi = C/\sqrt{T}$ where C is fixed (so that $\Pi'Z'Z\Pi$ converges to a constant as the sample size T grows), where $\hat{\beta}_{2SLS}$ is asymptotically biased. In finite samples $\hat{\beta}_{2SLS}$ is less biased than β_{OLS}, the OLS estimator of β in (5.5). The tails (and bias) of the finite sample distribution of $\hat{\beta}_{2SLS}$ depend on the degree of overidentification, the number of instruments excluded from the structural equation minus the dimension of β; the moments of the finite sample distribution exist up to/including this degree of overidentification.

In the method of limited information maximum likelihood, due to Anderson and Rubin (1949) and Hood and Koopmans (1953), the estimator for β is the value of β for which the likelihood function of (5.5)-(5.6), concentrated with respect to Π and Σ, takes its maximum. It is computed by computing the smallest root λ of the determinantal equation $|\lambda(Y\ \ X)'(Y\ \ X) - (Y\ \ X)'Z(Z'Z)^{-1}Z'(Y\ \ X)| = 0$ and the corresponding eigenvector, after which multiplying this eigenvector with minus the inverse of its first element yields $(-1, \hat{\beta}_{LIML})'$. Staiger and Stock (1997) show that in their case of weak instruments $\hat{\beta}_{LIML}$ is an inconsistent estimator of β. However, in finite samples $\hat{\beta}_{LIML}$ is (approximately) median unbiased if instruments are not too weak. Staiger and Stock (1997) show several cases in which the LIML estimator is approximately median unbiased whereas the 2SLS estimator suffers from huge biases, which makes Staiger and Stock (1997) conclude that estimator bias is less of a problem for LIML than for 2SLS, so that they suggest using LIML rather than 2SLS point estimates. The tails of the finite sample distribution of

$\hat{\beta}_{LIML}$ are Cauchy-type (no matter the degree of overidentification), so that $\hat{\beta}_{LIML}$ has no finite moments.

Angrist and Krueger (1991) report the 2SLS estimate of β in the model (5.5)-(5.6): $\hat{\beta}_{2SLS} = 0.0928$ with an asymptotic standard error of 0.0093 (column (2) of Table VII). Next to that they report the OLS estimate of β in (5.5): 0.0673 (with standard error of 0.0003; column (1) of Table VII).

Table 5.5 shows the results of 2SLS and OLS for the Census regions. This suggests that the 2SLS estimate for the US is almost completely determined by the region South: the difference between the 2SLS estimates for the US and the South is small, and the asymptotic standard error for the South is not much larger than that for the US. An explanation for this result is that the average education level for men born in 1930-1939 is lower in the region South than in the other regions, see Table 5.2. The influence of compulsory schooling laws is therefore larger for the South, as more students desire to leave school as soon as it is allowed. Therefore the influence of quarter of birth is larger in the Southern region, so that the instruments are strongest in the South.

One problem that the 2SLS estimator may suffer from is that it is biased in the case of weak instruments; this is illustrated by the last column of Table 5.5, which shows the mean of 10,000 2SLS estimators for 10,000 data sets simulated from (5.5)-(5.6) with parameter values chosen as $\beta = \hat{\beta}_{2SLS}$, $\Pi = \hat{\Pi}_{OLS}$ and Σ the covariance matrix of the residuals. The five means are all biased in the direction of the corresponding OLS estimator where the relative bias, the difference between the mean of the 2SLS estimates and the true β in the simulations divided by the difference between the OLS estimate and the true β, is smaller for the US and the Southern region than for the other three regions. In fact it is smaller for the South than for the US, which reflects that the addition of superfluous (or very weak) instruments results in a 2SLS estimator with smaller variance but larger bias.

Table 5.6 shows LIML estimates for the four Census regions, and quantiles of the estimated finite sample distribution, where maximum likelihood estimates substituted for β, Π and Σ in the finite sample density of $\hat{\beta}_{LIML}$ for the case of one explanatory endogenous variable that is given by Kleibergen (2000) and Kleibergen and Zivot (2003). Again the results are dominated by the Southern region: the difference between the LIML estimates for the US and the South is small, and the 95% and 50% density intervals for the South are not much larger than those for the US. Since LIML is known to focus on the strongest available instruments, this confirms that the instruments are much stronger

in the South than in the other regions. Also notice that the median of the finite sample distribution of the LIML estimator is approximately equal to the 'true value', the ML estimate, which reflects that the LIML estimator is approximately median unbiased. So, Tables 5.5 and 5.6 illustrate that the LIML estimator is a better point estimator than the 2SLS estimator, although in this case the 2SLS estimator is also 'smart enough' to indicate that the strongest instruments stem from the region South.

The results for the four Census regions suggest that a further division of the data set into states (or groups of states) may be interesting. We first take a closer look at the first stage regression, the OLS results in model (5.5). Table 5.7 shows the estimated coefficients $\hat{\pi}_{th}^{sq}$ at the interactions of state and quarter of birth, which are interpreted as the effect of the quarter of birth (as compared with the first quarter) in the year 1939. It shows t-values larger than 3 for Arkansas, Kentucky and Tennessee (and t-values larger than 2 for Alabama, Arizona, California, Colorado, Georgia, Illinois, Louisiana, Maryland, Massachusetss, Mississippi. North Carolina, North Dakota, Texas, Virginia). The effect of quarter of birth on education should be on average smaller than 0.75, which is not satisfied by the estimated coefficients of Alaska, Hawaii and Nevada; this is obviously caused by the small numbers of observations for these states. Table 5.8 shows the estimated coefficients $\hat{\pi}_{jh}^{yq}$ at the interactions of year and quarter of birth. This shows that the influence of quarter of birth is clearly strongest for men born in 1930. One explanation is that men born in 1930 have on average less education than men born in 1931-1939, so that compulsory schooling laws are more important for this group. Another, more specific, reason could be that these men attain age 16 in 1946, right after World War II when there is arguably a lot of work for young men.

Another way to look at the strength of the quarter of birth instruments for each state is to consider the p-value of the multiple F-test in the first stage regression when considering only data of one state. These F-statistics (and corresponding p-values) are given by Table 5.9, which also shows the estimated concentration parameter $\Pi' Z' Z \Pi / \sigma_{22}^2$ (with $\sigma_{22}^2 = \text{var}(v_i)$; see Basmann(1963)), where $\hat{\Pi}_{OLS}$ and the variance of the residuals in the first stage regression are substituted for Π and σ_{22}^2.

Table 5.9 shows that three states in the Census region South, Arkansas, Kentucky and Tennessee, have the largest concentration parameter, and the smallest p-values in the multiple F-test ($p < 0.001$). For Kansas the p-value in the multiple F-test is smaller than 0.01. We have $p < 0.1$ for Arizona and three Southern states, Georgia, South Carolina and Texas.

The results of the multiple F-test in the first stage regression for data of one state are graphically illustrated by Figure 5.1. The three states with p-value smaller than 0.001, Arkansas, Kentucky and Tennessee, are neighboring states in the region South.

Kentucky has the highest concentration parameter (and smallest p-value at the F-test); it is no coincidence that for men born in 1930-1939 those born in Kentucky have the lowest education on average, so that the influence of compulsory schooling laws is relatively large in Kentucky. Arkansas and Tennessee also have relatively low average education levels. However, Virginia and Mississippi have lower average education levels; for Tennessee the states of Alabama and West Virginia also have lower average education. This suggests that the average amount of education desired by people is not the only factor influencing the strength of the quarter of birth instruments: there are also other factors playing a role, which may include the power of government agencies enforcing schooling laws and the exemptions from these schooling laws that vary between states.

Tables 5.5 and 5.6 show the results of 2SLS and LIML for data of men born in Kentucky, Arkansas or Tennessee. Notice that the uncertainty in the LIML estimator, reflected by the 95% and 50% density intervals of the (estimated) finite sample distribution, increases by a relatively small amount, as compared with the US. The width of the 95% and 50% density intervals are only 1.93 and 1.92 times larger than for the US while the whole data set of the US has over 14 times as many observations (329509 vs. 23062). Further, these 95% and 50% density intervals are tighter for the data set of Kentucky, Arkansas and Tennessee than for the region Northeast, Midwest or West. This stresses the importance of the observations on men born in the states of Arkansas, Kentucky and Tennessee for the inference on return on education.

Table 5.5: OLS and 2SLS estimates for β in (5.5)-(5.6) for data of US and Census regions

		OLS		2SLS		2SLS (10000 simulations)	
	Region	$\hat{\beta}_{OLS}$	(st.error)	$\hat{\beta}_{2SLS}$	asympt. std.error	mean $\hat{\beta}_{2SLS}$	relative bias
	USA	0.0673	(0.0003)	0.0928	(0.0093)	0.0858	0.275
1	Northeast	0.0738	(0.0007)	0.0707	(0.0234)	0.0721	0.452
2	Midwest	0.0621	(0.0007)	0.0796	(0.0224)	0.0724	0.411
3	South	0.0691	(0.0006)	0.0931	(0.0120)	0.0874	0.238
4	West	0.0559	(0.0012)	0.0506	(0.0206)	0.0530	0.453
	Kentucky, Arkansas & Tennessee	0.0653	(0.0013)	0.0970	(0.0168)		

Table 5.6: LIML estimates for β in (5.5)-(5.6) for data of US and Census regions

			Quantile finite sample dist. $\hat{\beta}_{LIML}$				
	Region	$\hat{\beta}_{LIML}$	median	2.5%	97.5%	25%	75%
	USA	0.1064	0.106	0.0877	0.1256	0.0999	0.1129
1	Northeast	0.0650	0.065	0.0163	0.1124	0.0487	0.0810
2	Midwest	0.1298	0.128	0.0836	0.1825	0.1135	0.1468
3	South	0.1071	0.107	0.0828	0.1324	0.0986	0.1156
4	West	0.0449	0.045	0.0014	0.0874	0.0302	0.0593
	Kentucky, Arkansas & Tennessee	0.1046	0.104	0.0694	0.1420	0.0922	0.1170

Table 5.7: Estimated coefficients $\hat{\pi}_{th}^{sq}$ at interactions of state and quarter of birth in first stage regression

State	Second quarter		Third quarter		Fourth quarter	
	coefficient	(t-value)	coefficient	(t-value)	coefficient	(t-value)
Alabama	0.0087	(0.0799)	0.2339	(2.1891)	0.2591	(2.4017)
Alaska	2.0370	(1.9631)	1.6165	(1.4696)	-0.0412	(-0.0342)
Arizona	0.6195	(2.2070)	-0.1240	(-0.4569)	-0.2899	(-1.0466)
Arkansas	-0.2182	(-1.6787)	0.0271	(0.2175)	0.4230	(3.3300)
California	0.2272	(2.3296)	0.1102	(1.1609)	0.1029	(1.0606)
Colorado	0.2718	(1.5384)	0.3377	(1.9393)	0.3958	(2.2194)
Connecticut	0.2548	(1.6819)	0.0123	(0.0807)	0.0875	(0.5661)
Delaware	0.4562	(1.2199)	0.5355	(1.4456)	0.0847	(0.2232)
D.C.	-0.4522	(-1.7005)	-0.4764	(-1.8166)	-0.4318	(-1.6783)
Florida	0.2481	(1.6166)	0.1023	(0.6891)	0.2073	(1.4024)
Georgia	-0.2805	(-2.5634)	-0.0438	(-0.4112)	0.0417	(0.3827)
Hawaii	-0.0748	(-0.1191)	1.4881	(2.6138)	0.7564	(1.3884)
Idaho	0.1301	(0.5699)	-0.0874	(-0.3848)	0.1135	(0.4925)
Illinois	0.0294	(0.3604)	-0.1606	(-2.0227)	0.0464	(0.5712)
Indiana	-0.0111	(-0.1038)	-0.0385	(-0.3720)	0.0140	(0.1315)
Iowa	-0.0846	(-0.7066)	-0.1300	(-1.1160)	0.0574	(0.4814)
Kansas	0.2322	(1.6350)	0.2251	(1.6784)	0.1554	(1.1369)
Kentucky	0.0492	(0.4631)	0.2550	(2.4374)	0.5142	(4.8730)
Louisiana	0.0434	(0.3328)	0.1172	(0.9531)	0.2686	(2.1695)
Maine	-0.0241	(-0.1276)	0.1437	(0.7740)	0.0746	(0.3942)
Maryland	0.3283	(2.2095)	0.3100	(2.1400)	0.2962	(2.0239)
Massachusetts	0.0894	(0.8809)	0.0678	(0.6789)	0.2404	(2.3322)
Michigan	0.1278	(1.4397)	0.0133	(0.1521)	0.1005	(1.1230)
Minnesota	-0.2100	(-1.8148)	-0.2733	(-2.3791)	-0.1192	(-1.0244)
Mississippi	0.0230	(0.1813)	0.1199	(0.9743)	0.3059	(2.4262)
Missouri	-0.1274	(-1.2029)	-0.0186	(-0.1827)	-0.0304	(-0.2922)
Montana	-0.0333	(-0.1376)	0.0058	(0.0234)	0.3234	(1.3071)
Nebraska	-0.1431	(-0.8979)	-0.2160	(-1.3662)	-0.1432	(-0.8997)
Nevada	-0.1005	(-0.1757)	-0.0600	(-0.1113)	0.7544	(1.3369)
New Hampshire	-0.1475	(-0.5482)	-0.0027	(-0.0104)	0.1464	(0.5420)
New Jersey	0.0372	(0.3524)	-0.0448	(-0.4308)	0.1848	(1.7262)
New Mexico	0.1925	(0.7601)	0.0274	(0.1052)	0.4056	(1.6175)
New York	0.0571	(0.8143)	-0.0631	(-0.9140)	-0.0776	(-1.1021)
North Carolina	-0.0857	(-0.8653)	0.0385	(0.3969)	0.2136	(2.1662)
North Dakota	-0.4979	(-2.4204)	-0.3019	(-1.4854)	-0.1407	(-0.6779)

Table 5.7 (continued)

State	Second quarter		Third quarter		Fourth quarter	
	coefficient	(t-value)	coefficient	(t-value)	coefficient	(t-value)
Ohio	-0.0847	(-1.0210)	-0.0583	(-0.7150)	0.0232	(0.2794)
Oklahoma	-0.0800	(-0.6610)	0.0400	(0.3501)	0.2066	(1.7793)
Oregon	0.0550	(0.2705)	-0.0238	(-0.1208)	0.0233	(0.1145)
Pennsylvania	-0.0288	(-0.3996)	-0.0959	(-1.3526)	0.0174	(0.2408)
Rhode Island	-0.3987	(-1.8053)	0.0585	(0.2624)	0.1230	(0.5475)
South Carolina	-0.1087	(-0.8025)	-0.0978	(-0.7348)	0.2983	(2.2239)
South Dakota	0.2971	(1.3435)	0.1932	(0.8849)	0.5111	(2.2999)
Tennessee	-0.1076	(-0.9929)	0.0848	(0.7879)	0.4465	(4.0963)
Texas	-0.0935	(-1.0690)	0.1063	(1.2806)	0.2505	(2.9811)
Utah	-0.0264	(-0.1254)	-0.2249	(-1.0941)	-0.2249	(-1.0827)
Vermont	0.1709	(0.5889)	0.3196	(1.0907)	0.2557	(0.8531)
Virginia	0.0300	(0.2608)	0.1905	(1.6867)	0.2881	(2.4945)
Washington	0.1001	(0.6368)	0.0280	(0.1802)	0.0009	(0.0054)
West Virginia	-0.0901	(-0.7396)	-0.0093	(-0.0774)	0.2320	(1.9120)
Wisconsin	0.1516	(1.4148)	-0.1032	(-0.9689)	0.0702	(0.6479)
Wyoming	0.0806	(0.2302)	0.1759	(0.5205)	-0.1905	(-0.5308)

Table 5.8: Estimated coefficients $\hat{\pi}_{jh}^{yq}$ at interactions of quarter and year of birth in first stage regression (1939 = reference year)

State	Second quarter		Third quarter		Fourth quarter	
	coefficient	(t-value)	coefficient	(t-value)	coefficient	(t-value)
1930	0.1538	(2.2272)	0.1881	(2.7767)	0.2191	(3.1690)
1931	-0.0378	(-0.5337)	0.1470	(2.1176)	-0.0472	(-0.6650)
1932	0.0804	(1.1494)	0.0977	(1.4301)	0.0962	(1.3841)
1933	-0.0677	(-0.9584)	0.0681	(0.9805)	-0.0989	(-1.4070)
1934	0.0792	(1.1212)	0.0574	(0.8356)	0.0340	(0.4865)
1935	0.1162	(1.6620)	0.1856	(2.7248)	0.0170	(0.2433)
1936	0.0274	(0.3931)	0.1067	(1.5671)	0.0209	(0.3009)
1937	0.0120	(0.1729)	0.1372	(2.0335)	0.0472	(0.6843)
1938	0.0396	(0.5765)	0.0291	(0.4354)	-0.0348	(-0.5096)

Table 5.9: Summary statistics of first-stage regression for data of individual states

State	concentration parameter	R^2	F-stat.	# obs.	p-value F-stat.
Alabama	18.68	0.0022	0.62	8536	0.9463
Arizona	45.64	0.0426	1.52	1066	0.0365
Arkansas	60.06	0.0103	2.00	5794	0.0009
California	35.93	0.0032	1.20	11078	0.2107
Colorado	31.68	0.0113	1.06	2818	0.3836
Connecticut	38.25	0.0100	1.27	3844	0.1448
Delaware	29.74	0.0506	0.99	598	0.4814
D.C.	27.64	0.0226	0.92	1237	0.5892
Florida	36.19	0.0093	1.21	3913	0.2031
Georgia	41.29	0.0049	1.38	8411	0.0827
Hawaii	27.70	0.1185	0.92	246	0.5853
Idaho	22.57	0.0143	0.75	1599	0.8309
Illinois	32.74	0.0018	1.09	18375	0.3341
Indiana	27.12	0.0030	0.90	8918	0.6169
Iowa	25.64	0.0038	0.85	6699	0.6931
Kansas	52.35	0.0109	1.75	4807	0.0072
Kentucky	68.54	0.0076	2.28	8933	0.0001
Louisiana	31.18	0.0052	1.04	5975	0.4071
Maine	24.15	0.0100	0.80	2424	0.7645
Maryland	34.10	0.0083	1.14	4139	0.2777
Massachusetts	36.02	0.0036	1.20	9955	0.2080
Michigan	23.78	0.0017	0.79	14077	0.7818
Minnesota	15.82	0.0022	0.53	7170	0.9841
Mississippi	30.43	0.0052	1.01	5864	0.4442
Missouri	28.46	0.0031	0.95	9274	0.5460
Montana	26.99	0.0194	0.90	1407	0.6233
Nebraska	30.94	0.0089	1.03	3488	0.4189
Nevada	25.42	0.0866	0.85	308	0.6988
New Hampshire	25.18	0.0212	0.84	1200	0.7147
New Jersey	31.73	0.0035	1.06	8964	0.3803
New Mexico	26.10	0.0199	0.87	1325	0.6690
New York	38.23	0.0013	1.27	29015	0.1440
North Carolina	40.10	0.0037	1.34	10798	0.1033
North Dakota	30.95	0.0153	1.03	2028	0.4191

Table 5.9 (continued)

State	concentration parameter	R^2	F-stat.	# obs.	p-value F-stat.
Ohio	32.16	0.0019	1.07	17070	0.3601
Oklahoma	36.85	0.0053	1.23	6950	0.1822
Oregon	20.37	0.0097	0.68	2127	0.9054
Pennsylvania	27.70	0.0011	0.92	26385	0.5862
Rhode Island	33.96	0.0201	1.13	1698	0.2850
South Carolina	50.91	0.0097	1.70	5245	0.0102
South Dakota	38.39	0.0219	1.28	1754	0.1428
Tennessee	64.06	0.0077	2.14	8335	0.0003
Texas	50.01	0.0031	1.67	15932	0.0125
Utah	31.85	0.0160	1.06	1999	0.3761
Vermont	36.54	0.0367	1.22	999	0.1960
Virginia	28.61	0.0039	0.95	7319	0.5384
Washington	18.54	0.0052	0.62	3610	0.9487
West Virginia	34.36	0.0054	1.15	6412	0.2673
Wisconsin	27.27	0.0032	0.91	8607	0.6090
Wyoming	29.27	0.0421	0.98	706	0.5050

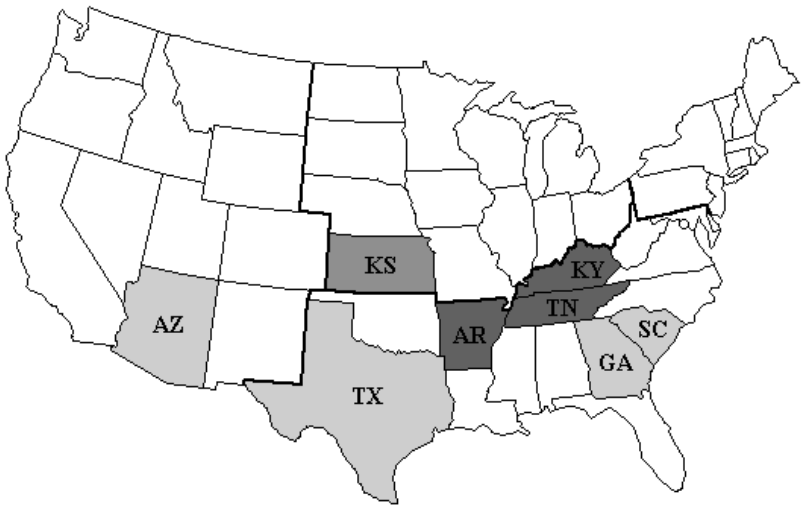

Figure 5.1: p-value of multiple F-test in first stage regression for data of individual states: p-value < 0.001*: dark grey, p-value* < 0.01*: grey, p-value* < 0.1*: light grey.*
(AR = Arkansas, AZ = Arizona, GA = Georgia, KS = Kansas, KY = Kentucky, SC = South Carolina, TN = Tennessee, TX = Texas)

5.4 Bayesian approaches

In this section we first briefly discuss the posterior distributions under two commonly used prior density kernels, the flat prior of Drèze (1976) and the Jeffreys prior. For an extensive discussion of these Bayesian approaches (and their relations to the 2SLS and LIML estimators) the reader is referred to Kleibergen and Zivot (2003). After that the results are discussed of applying these Bayesian methods to the Angrist-Krueger IV model for data of the US and the four Census regions.

Drèze (1976) specifies the following flat prior on the parameters of the structural form (5.5)-(5.6):

$$p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-1/2\,(k+m+2)} \tag{5.9}$$

where k is the number of instruments in Z and m is the number of explanatory endogenous variables in X. The primary motivation of this flat prior is that it has an invariance property in the sense that the prior on the structural form implies the same kind of prior on the parameters of the restricted reduced form (which is proportional to $|\Sigma|^{-1/2\,(k+m+2)}$). The marginal posterior of β resulting from the prior (5.9) is given by:

$$p(\beta|y, X, Z) \propto \left(\frac{(y - X\beta)' M_Z (y - X\beta)}{(y - X\beta)'(y - X\beta)}\right)^{T/2} (y - X\beta)'(y - X\beta)^{-k/2}, \tag{5.10}$$

where $M_Z \equiv I - Z(Z'Z)^{-1}Z'$. The tails of this posterior of β become thinner when (possibly superfluous) instruments are added to the model, see *e.g.* Maddala (1976) and Kleibergen and Zivot (2003). Further, the location of the posterior mode moves towards the OLS estimator when superfluous instruments are added. Bayesian inference under the flat prior of Drèze (1976) shares these properties with the small sample distribution of the 2SLS estimator which made Kleibergen and Zivot (2003) conclude that this approach has more in common with 2SLS than with LIML.

The Jeffreys prior, the square root of the determinant of the information matrix, is given by:

$$p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-(m+1)} |\Pi' Z' Z \Pi|^{1/2} |\Sigma_{22.1}|^{-1/2(k-m)} \tag{5.11}$$

with $\Sigma_{22.1} \equiv \Sigma_{22} - \Sigma_{21}\sigma_{11}^{-1}\Sigma_{12}$ for the structural form (5.5)-(5.6) or equivalently by:

$$p(\beta, \Pi, \Omega) \propto |\Omega|^{-(m+1)} |\Pi' Z' Z \Pi|^{1/2} |(\beta : I_m)\Omega^{-1}(\beta : I_m)'|^{1/2(k-m)} \tag{5.12}$$

for the corresponding restricted reduced form (5.7)-(5.8); see for example Appendix A of Hoogerheide, Kleibergen and Van Dijk (2006) for a derivation of this Jeffreys prior. In

the case of $m = 1$ and for moderate values of T ($T > 20$), an accurate approximation of the marginal posterior of β can be obtained by

$$p(\beta|\Omega, y, X, Z) \propto [(\beta - \phi)^2\omega_{11.2}^{-1} + \omega_{22}^{-1}]^{-1} \times \sum_{j=0}^{\infty} \frac{\Gamma[(k+2j+1)/2]}{j!\,\Gamma[(k+2j)/2]} \left(\frac{(\beta:1)\Omega^{-1}\hat{\Phi}'Z'Z\hat{\Phi}\Omega^{-1}(\beta:1)'}{2[(\beta-\phi)^2\omega_{11.2}^{-1} + \omega_{22}^{-1}]} \right)^j \quad (5.13)$$

where $\Omega = (y : X)'(y : X)/T$ is substituted for Ω, and where $\phi \equiv \omega_{21}/\omega_{22}$, $\omega_{11.2} \equiv \omega_{11} - \omega_{21}^2/\omega_{22}$, $\hat{\Phi} = (Z'Z)^{-1}Z'(y : X)$, see Kleibergen and Zivot (2003). The primary motivation of the Jeffreys prior is its universal invariance property with respect to parameter transformations. Kleibergen and Zivot (2003) show that Bayesian analysis using a Jeffreys prior leads to, when there is only $m = 1$ explanatory endogenous variable, a functional expression of the marginal posterior of β that is identical to the finite sample density of the LIML estimator. Just like the finite sample distribution of the LIML estimator, the posterior based on the Jeffreys prior retains Cauchy type tails when (possibly irrelevant) instruments are added, and the location of the mode is insensitive to the addition of superfluous instruments.

Table 5.10 shows some summary statistics of the posterior distribution of β under the flat or Jeffreys prior. Just like the results for the 2SLS and LIML estimators, the posterior distribution of β for the US under the flat or Jeffreys prior is almost completely determined by the region South; the difference between the means or medians for the US and the South is small, and the posterior standard deviation and 95% and 50% posterior density intervals for the South are not much larger than those for the US, whereas the posterior density intervals are relatively large for the other regions.[6] It is shown in Hoogerheide, Kleibergen and Van Dijk (2006) that Bayesian analysis using the Jeffreys prior, similar to the LIML estimator, focusses on the strongest available instruments. So, the posterior results under the Jeffreys prior once more indicate that the quarter of birth instruments are strongest in the South. Figure 5.2 shows the graphs of the posterior densities under the flat and Jeffreys prior, respectively.

For all four approaches that have been considered in this chapter, the two classical methods as well as the two Bayesian approaches, inference on the return to education for

[6]These 95% and 50% posterior density intervals are not equal to 95% and 50% Highest Posterior Density (HPD) regions, although the differences are small in these cases of unimodal, almost symmetric distributions.

the US is almost completely determined by the returns to education in the South. If the effect of the return on education is different for the other regions, which can not a priori be ruled out given the large economic differences between these regions, inference using data of the US is not representative for the average returns on education across the US. One should thus be careful when drawing such conclusions.

Notice that the results for the flat prior are remarkably similar to those for the 2SLS estimator: for each region the posterior mean of β is close to the 2SLS estimator $\hat{\beta}_{2SLS}$ and the posterior standard deviation is close to the asymptotic standard error. This agrees with the conclusions of Kleibergen and Zivot (2003) that Bayesian analysis using the flat prior is closer to 2SLS than to LIML. Also note that the results for the Jeffreys prior are similar to those for the LIML estimator: the posterior median of β is close to the LIML estimator $\hat{\beta}_{LIML}$, and for US and the region South the difference in the 95% and 50% intervals between the methods is close, although for the other three Census regions the 95% and 50% posterior intervals under the Jeffreys prior are somewhat larger than the corresponding intervals for the (estimated) finite sample distribution of the LIML estimator.

We also consider the posterior of β under the flat and Jeffreys prior based on only the observations on men born in the states of Arkansas, Kentucky and Tennessee. Table 5.10 shows summary statistics of these posteriors. Notice that the uncertainty in the posterior under the Jeffreys prior, reflected by the 95% and 50% density intervals of the (estimated) finite sample distribution, increases by a relatively small amount, as compared with the US. The width of the 95% and 50% posterior density intervals are only 1.63 and 1.56 times larger than for the US while the whole data set of the US has over 14 times as many observations (329509 vs. 23062). Further, these 95% and 50% posterior density intervals are tighter for the data set of Kentucky, Arkansas and Tennessee than for the region Northeast, Midwest or West. Figure 5.3 illustrates the relative strength of the quarter of birth instruments in the states of Arkansas, Kentucky and Tennessee. If we divide the data set of the US in three subsamples, Arkansas-Kentucky-Tennessee (23062 observations), the other 14 states of region South (91329 observations) and the other three Census regions (215118 observations), then the resulting posteriors of β under the Jeffreys prior are about as tight for these three subsamples. These results again stress the importance of the states of Arkansas, Kentucky and Tennessee for the inference on return on education: to a large extent inference on education for the US is determined by the return on education for men born in these three states.

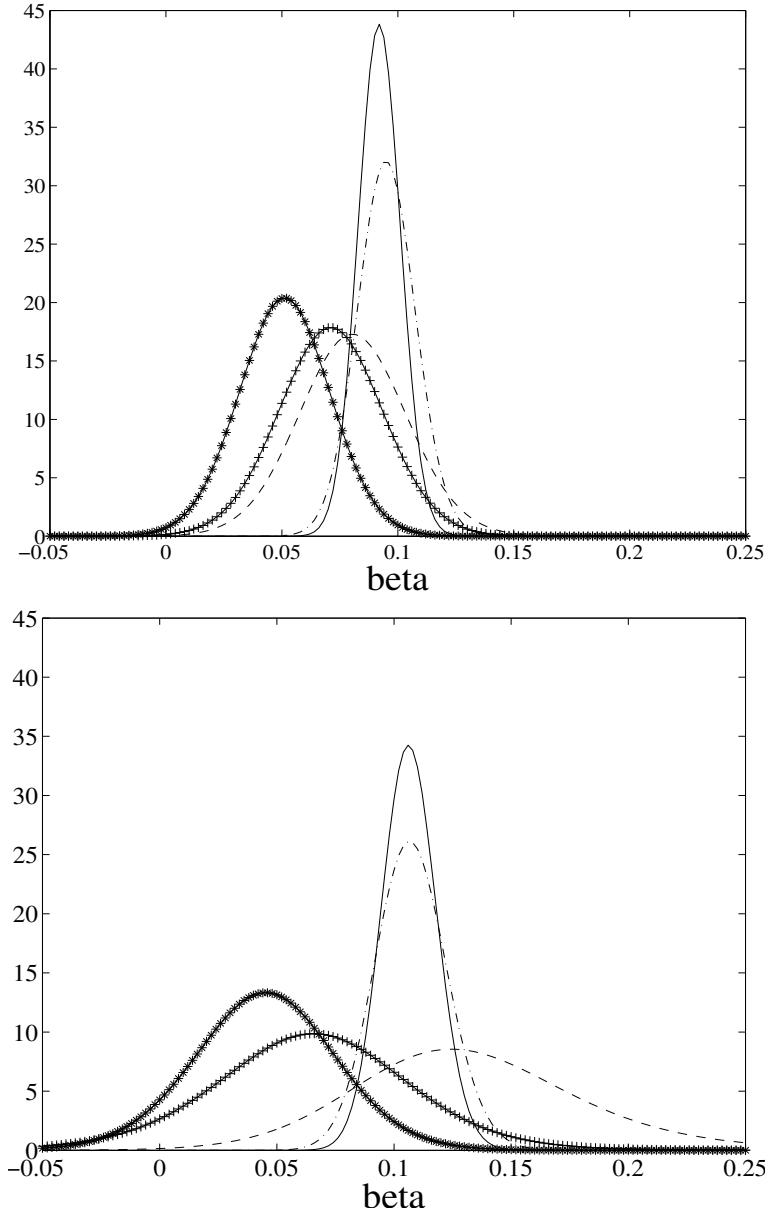

Figure 5.2: Marginal posterior of return on education β under flat prior (above) or Jeffreys prior (below) for US (solid), Northeast (solid-plusses), Midwest (dashed), South (dash-dot), West (solid with stars).

Table 5.10: Posterior results under flat or Jeffreys prior for US and regions

		Posterior β under flat prior		Quantile posterior β under Jeffreys prior				
	Region	mean	st.dev.	median	2.5%	97.5%	25%	75%
	USA	0.092	0.009	0.106	0.083	0.129	0.098	0.114
1	Northeast	0.071	0.023	0.064	-0.024	0.150	0.037	0.092
2	Midwest	0.081	0.023	0.129	0.041	0.246	0.099	0.163
3	South	0.095	0.012	0.107	0.077	0.138	0.096	0.117
4	West	0.051	0.020	0.044	-0.018	0.105	0.024	0.065
	Kentucky, Arkansas & Tennessee	0.095	0.016	0.104	0.068	0.143	0.092	0.117

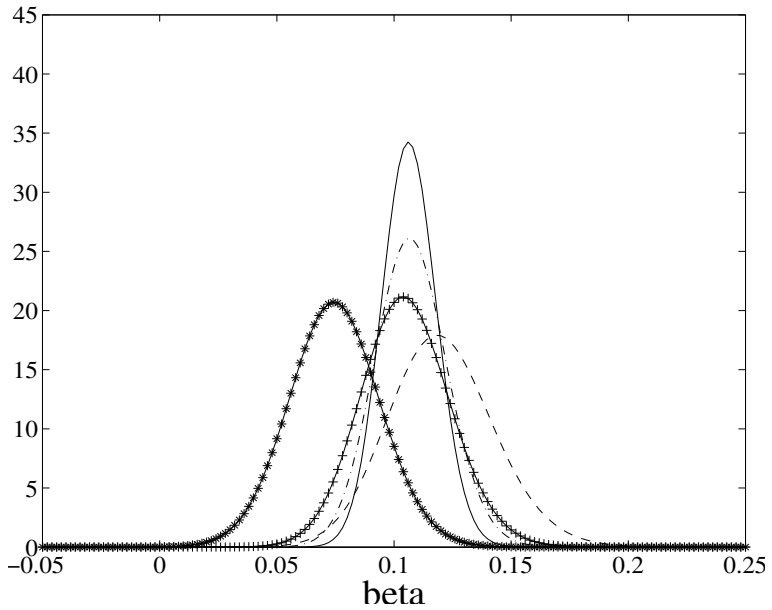

Figure 5.3: Marginal posterior of return on education β under Jeffreys prior for US (solid), region South (17 states, dash-dot), Kentucky-Tennessee-Arkansas (solid with plusses), rest of region South (14 states, dashed), other three Census regions (34 states, solid with stars).

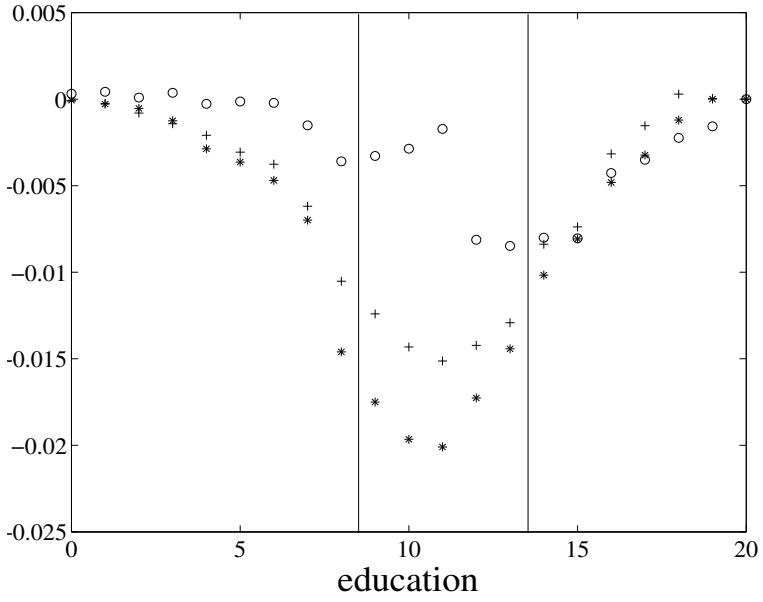

Figure 5.4: Empirical cumulative distribution function (CDF) of (completed years of) education for men born in different quarters: difference between CDF for men born in second quarter (circle)/ third quarter (plus)/ fourth quarter (star) and CDF for men born in first quarter.

5.5 Investigation of some of the assumptions made by Angrist and Krueger (1991)

Angrist and Krueger (1991) make the assumption that the only reason for the influence of quarter of birth on education is the asymmetry between the school-entry requirements and compulsory schooling laws: a child's birthday determines whether the school district allows the child to enter school at age 5 or age 6, whereas compulsory schooling laws generally allow students to immediately leave school when they reach a certain age (mostly 16, sometimes 17 or 18). Reasoning in this way, the quarter of birth should only yield valuable instruments for education for those individuals who have completed 9 - 13 years of education, as all persons who have left school as soon as the law allowed for it should be contained in this group.

We first inspect the empirical cumulative distribution function (CDF) of years of education for the four quarters of birth. If quarter of birth would only affect the education spell for those who leave school as soon as the law allows for it, the CDF of education should only differ for the range $9 \leq \text{education} \leq 13$. Figure 5.4 shows the difference

Table 5.11: First stage regressions for subsamples of the region South: R^2, F-statistic of multiple F-test and corresponding p-value

	R^2	F-statistic	obs.	p-value
region South	0.0023	3.3333	114391	0.0000
education ≤ 8	0.0045	1.1034	19214	0.2495
$9 \leq$ education ≤ 13	0.0019	1.5519	63637	0.0013
education ≥ 14	0.0029	1.1790	31540	0.1339
education ≤ 8 or education ≥ 14	0.0043	2.8060	50754	0.0000

between the empirical CDF of education between quarter 2/3/4 and quarter 1. This shows that also for education ≤ 8 and for education ≥ 14 the CDF substantially differs between the quarters of birth. Notice the negative values reflecting that men born in the first quarter have completed less years of education on average (also conditional on education ≤ 8 or education ≥ 14).

In order to investigate the importance of the observations with education ≤ 8 or education ≥ 14 for the results in the IV model we now divide the data set of the Southern region (that determines the results for the US) into a subsample of men with 9 - 13 years of education and subsamples of men with at most 8 years or at least 14 years of education. Figure 5.5 shows the number of observations per years of education in the region South. Table 5.11 shows the R^2 and F-statistic of the multiple F-test in the first stage regression for several subsamples of the region South based on education levels. Notice that the R^2 is even lower for $9 \leq$ education ≤ 13 than for the groups with education ≤ 8 and education ≥ 14, suggesting quarter of birth instruments are even stronger for people with education outside the interval 9 - 13 than inside this interval. If we look at the group with either education ≤ 8 or education ≥ 14, which consists of a number of observations comparable to the group with $9 \leq$ education ≤ 13, then we see that the p-value at the multiple F-statistic is also smaller for men with years of education outside the interval 9 - 13. This suggests that the influence of compulsory schooling laws on students who want to leave school as early as possible is certainly not the only factor causing the effect of quarter of birth on (average) education spell.

This is further illustrated by the posterior of β under the Jeffreys prior in Figure 5.6. The posterior under the Jeffreys prior is tighter for observations with education outside

the interval 9-13 than inside this interval. In fact, this posterior is even tighter than for the region South or the US; intuitively speaking, this is possible since not only one's knowledge on β is updated by the extra observations, but also on Π (which occurs as the product $\Pi\beta$ in the restricted reduced form of the model) and Ω. That the posterior is much tighter for the group with less than 9 or more than 13 years of education and that the posterior for all observations of the South is much closer to this posterior than to the posterior for data on men with $9 \leq$ education ≤ 13, suggests that instruments are truly (much) stronger for men with education ≤ 8 and education ≥ 14. For completeness, Figure 5.7 shows the posterior under the flat prior, which shows approximately the same shapes. These results suggest that quarter of birth does not only affect years of schooling for those who leave school as soon as it is allowed.

A possible explanation is that the probability that a student leaves school during a quarter depends not only on the number of quarters of schooling that the student has already had, but also (positively) on age (measured in quarters of years): children born in the first quarter enter school at a later age (measured in quarters), so that in each cohort the students born in the first quarter are the oldest. Reasoning in this way, the influence of quarter of birth on age at school entry is enough to cause exogenous variation in years of education, even without requiring laws keeping (a certain percentage of) students at school until they reach a certain age. In other words, quarter of birth influences the age at school entry, so that it causes an exogenous variation in education level as long as students with different ages (with age measured in quarters of years) have a different 'hazard rate' of quitting school after a certain amount of education. So, the results suggesting that the influence of quarter of birth on education is certainly not restricted to men who have completed 9-13 years of education does not imply that the model is useless. It only suggests that the strength of the quarter of birth instruments is not so much caused by the asymmetry between school entry requirements and compulsory schooling laws keeping students at school until they reach a certain age; the value of the quarter of birth instruments seems to stem to a larger extent from the school entry requirements in combination with the dependence of the 'hazard rate' of leaving school on age (measured in quarters).

Bound, Jaeger and Baker (1995) criticize the assumptions of Angrist and Krueger (1991). They draw attention to two problems associated with the use of the 2SLS estimator in the case of weak instruments. First, the use of weak instruments may lead to large

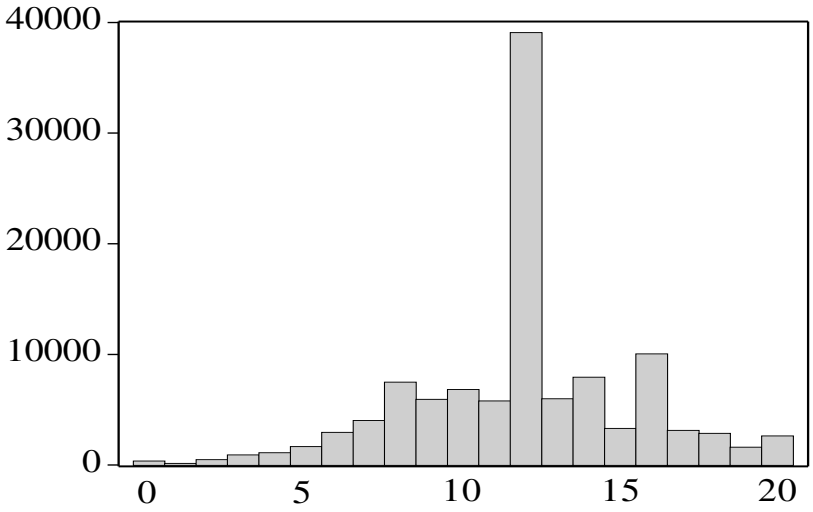

Figure 5.5: Histogram of the number of completed years of education in US Census region South

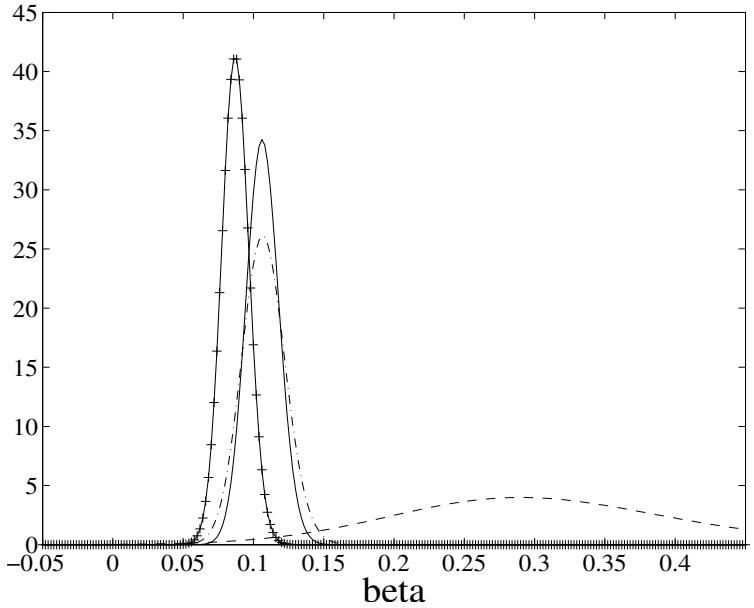

Figure 5.6: Marginal posterior of return on education β under Jeffreys prior for US (solid), South (17 states,dash-dot), South for education ≤ 8 or education ≥ 14 (solid-plusses), South for $9 \leq$ education ≤ 13 (dashed).

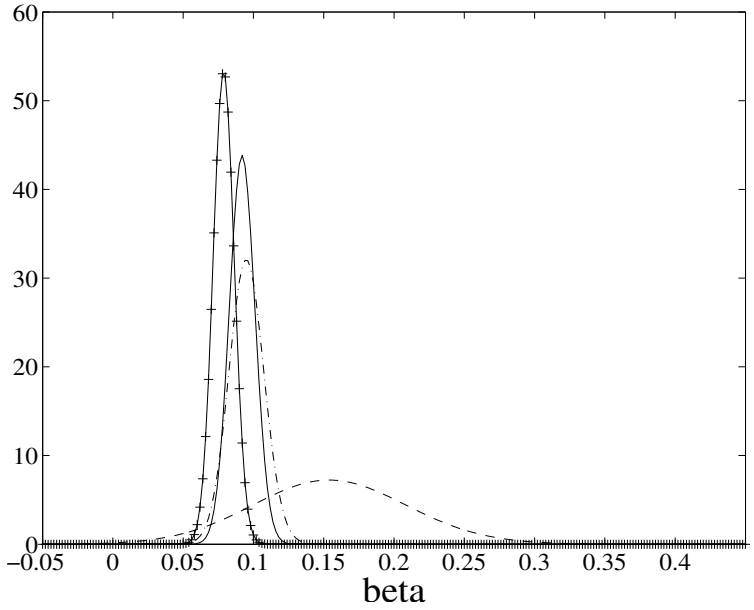

Figure 5.7: Marginal posterior of return on education β under flat prior for US (solid), South (17 states,dash-dot), South for education ≤ 8 or education ≥ 14 (solid-plusses), South for $9 \leq$ education ≤ 13 (dashed).

inconsistencies in the 2SLS estimator even if there is only a weak relationship between the instruments and the error in the structural equation.[7] Second, in finite samples, the 2SLS estimator is biased in the same direction as the OLS estimator, where the bias of the 2SLS estimator approaches that of the OLS estimator as the R^2 between instruments and explanatory endogenous variable approaches 0.

We first consider the problem of the finite-sample bias. Bound, Jaeger and Baker (1995) show that the 2SLS estimators of Angrist and Krueger (1991) may suffer from substantial finite-sample bias even with the large sample size, because the correlation between quarter of birth and years of education is only small. The simulations in table 5.5 confirm this; the 2SLS estimators for the US and the Census regions seem to have biases between 0.2 and 0.5 times the bias of the OLS estimator. However, the LIML estimator seems to be approximately median unbiased in these cases, suggesting that the

[7]Bound, Jaeger and Baker (1995) consider the case of weak instruments in the sense of instruments explaining little of the variation in the endogenous explanatory variable(s); this differs from the weak instrument defined as $\Pi = C/\sqrt{T}$ where C is fixed (so that $\Pi' Z' Z \Pi$ converges to a constant as the sample size T grows) considered by Staiger and Stock (1997).

problem of the finite-sample bias caused by the weakness of the instruments could be solved by using LIML instead of 2SLS in order to obtain a point estimate of β.

We now consider the problem of large inconsistencies even if there is only a weak relationship between the instruments and the error in the structural equation. Bound, Jaeger and Baker (1995) argue that a weak correlation between quarter of birth and wage (independent of the effect of quarter of birth on education) exists and that this correlation is large enough to have substantial effects on the estimates of Angrist and Krueger (1991). Bound, Jaeger and Baker (1995) mention several publications containing evidence suggesting that quarter of birth directly influences wages for four reasons: there is some evidence that (1) quarter of birth influences a student's performance at school, for example performance in reading, writing and arithmetic; (2) quarter of birth affects the probability that an individual will suffer from certain mental or physical diseases/disabilities such as schizophrenia, multiple sclerosis, manic depression and dyslexia; (3) there are regional patterns in birth seasonality; (4) children born in families with high incomes are are less likely to be born in winter months. Therefore, Bound, Jaeger and Baker (1995) conclude that it is questionable whether the assumption of no direct effect of quarter of birth on income is justified.

At points (2) it should be noted that men with no income in 1979 are excluded from the data set of Angrist and Krueger (1991), so that some of the men suffering from the mentioned diseases/disabilities may be excluded from the data set. At point (3) it should be noticed that part of the regional patterns in birth seasonality are 'filtered' by the state dummies that are included in the wage equation. However, these two factors obviously do not take away the doubt on the assumption of no direct effect of quarter of birth on income.

This doubt and the finite-sample bias of the 2SLS estimator made Bound, Jaeger and Baker (1995) conclude that *'the "natural experiment" afforded by the interaction between compulsory school attendance laws and quarter of birth does not give much usable information concerning the causal effect of education on earnings'*.

Bound, Jaeger and Baker (1995) even report that differences in family income at time of birth (point (4)) would seem to account for virtually all of the association between quarter of birth and wages, which results from the following reasoning. It is argued that the difference in mean log per capita income between those born in the first quarter and the others is at least -0.0238, as this difference of -0.0238 is observed for men born in more recent years and the seasonal variation in fertility has declined since the 1930s (in which the men of our data set are born). Further, Solon (1992) and Zimmerman (1992)

both found an intergenerational correlation in long-run income of at least 0.4, so that men born in the first quarter are expected to earn about 0.95% more than those born during the rest of the year. Bound, Jaeger and Baker (1995) report that for men born during the 1930s, those born in the first quarter earn 1.1% lower wages on average, hardly more than the 0.95% resulting from differences in family income at birth.

We now take a closer look at this result that differences in family income at birth between those born in the first quarter and those born during rest of the year would seem to explain all of the effect of quarter of birth on income. First, for men born in the region South (who determine the results for the US of both the classical and Bayesian methods used in this chapter) the difference between those born in first quarter and the rest is higher: 1.65% (measured as difference in mean log income). For men born in Arkansas, Kentucky, Tennessee (who have a substantial influence on the results within the region South) this difference is even 1.99%. Of course, the difference in family income at birth and/or the intergenerational correlation may also be larger for these regions; this is a topic for research.

Second, the phenomenon that those born in rich families are less likely to be born in winter months can be modelled by including a dummy variable indicating whether a person is born in the first quarter in the wage equation of the model (and dropping one of the interactions of state and quarter dummies from the set of instruments). We consider this model for observations on men born in Arkansas, Kentucky or Tennessee.[8] In the second stage regression of 2SLS the first quarter dummy has an insignificant (and even positive) estimated coefficient of 0.0028 (with standard error 0.0111). The results of 2SLS and LIML are given by Tables 5.12 and 5.13, where the quantiles of the finite-sample distribution of the LIML estimator are estimated by substituting the ML estimates into the formula for this finite-sample density in Kleibergen (2000) and Kleibergen and Zivot (2003). Table 5.14 gives the results of Bayesian inference under the flat and Jeffreys prior. For the Jeffreys prior Figure 5.8 shows the marginal posterior of β. This shows that the uncertainty in the classical estimators and the posteriors under a flat or Jeffreys prior

[8]There are two reasons for confining ourselves to data of Arkansas, Kentucky and Tennessee, the three states for which the quarter-of-birth instruments are strongest, in this case. First, the addition of extra variables in the wage equation substantially increases problems of multicollinearity, which are smaller when considering less states and hence less state-of-birth dummies and interacted state-and-quarter-of-birth dummies. Second, the assumption that if a direct effect of quarter of birth on wages exists, that this effect is constant across states, seems to be more realistic for a region of three neighboring states than for other (sub)samples.

does not increase much by including a first quarter dummy in the wage equation. In other words, a rather tight posterior for β is obtained using quarter-of-birth information, even if we drop assumption of no influence of first quarter on income.

We can also go somewhat further in the sense of including three quarter-of-birth dummies in the wage equation, so that not only a difference between the first quarter and quarters 2-4 is allowed, but differences are permitted between all four quarters. Tables 5.12, 5.13 and 5.14, and Figure 5.8 also show the results for this model. The inclusion of two more dummies clearly increases the uncertainty in the estimators and posteriors for β as the instruments are weaker in this case. Intuitively, this can be explained as follows: in this adapted model the strength of the instruments depends on the variation between the effects of quarters of birth across states (and years) instead of the size of these effects. For example, if the effect of quarter 2-4 versus quarter 1 is substantial but (approximately) the same for all states, then the instruments Z (residuals in the regression of $\tilde{Z}$ on W) will be (almost) superfluous in this adapted model, while the instruments may be rather strong in the original model.

Including quarter-of-birth dummies in the wage equation may result in (much) wider posterior intervals. Next to that, if a direct effect of quarter of birth on income exists, this may not be constant across states/years; in that case more terms should be added to the wage equation. So, an important question remains if such terms should be included in the wage equation and if so, how these should be specified. We leave this as a topic for further research.

Still, it should at least be noted that if there exists a direct effect of quarter of birth on income, it is not likely that the factors causing this effect differ between states/years in the same way as compulsory schooling laws and the degree to which these are enforced. So, even if there exists a direct effect of quarter of birth on income which varies across states/years, the difference between these effects and the effect of compulsory schooling laws may be exploited, so that the model may still give usable information on the causal effect of education on income.

Notice that since we can use the (approximately median unbiased) LIML estimator instead of the (biased) 2SLS estimator, and since we may still obtain a rather tight posterior for β if we allow for a direct effect of birth during the first quarter on income, it seems that the conclusion of Bound, Jaeger and Baker (1995) that the interaction between compulsory school attendance laws and quarter of birth does not give much usable information concerning the causal effect of education on earnings may have been too strong.

Table 5.12: 2SLS estimates for β for data of Kentucky, Arkansas and Tennessee

	2SLS	
Model	$\hat{\beta}_{2SLS}$	asympt. std.error
original model	0.0970	(0.0168)
+ dummy for first quarter of birth in wage equation	0.0986	(0.0182)
+ 3 dummies for quarters of birth in wage equation	0.0928	(0.0274)

Table 5.13: LIML estimates for β for data of Kentucky, Arkansas and Tennessee

		Quantile finite sample dist. $\hat{\beta}_{LIML}$				
Model	$\hat{\beta}_{LIML}$	median	2.5%	97.5%	25%	75%
original model	0.105	0.104	0.069	0.142	0.092	0.117
+ 1st quarter dummy	0.109	0.108	0.070	0.149	0.095	0.121
+ 3 quarter dummies	0.121	0.121	0.065	0.187	0.101	0.141

Table 5.14: Posterior results under flat or Jeffreys prior for data of Kentucky, Arkansas and Tennessee

	Posterior β under flat prior		Quantile posterior β under Jeffreys prior				
Model	mean	st.dev.	median	2.5%	97.5%	25%	75%
original model	0.095	0.016	0.104	0.068	0.143	0.092	0.117
+ 1st quarter dummy	0.097	0.018	0.108	0.068	0.152	0.094	0.122
+ 3 quarter dummies	0.090	0.026	0.121	0.043	0.220	0.094	0.150

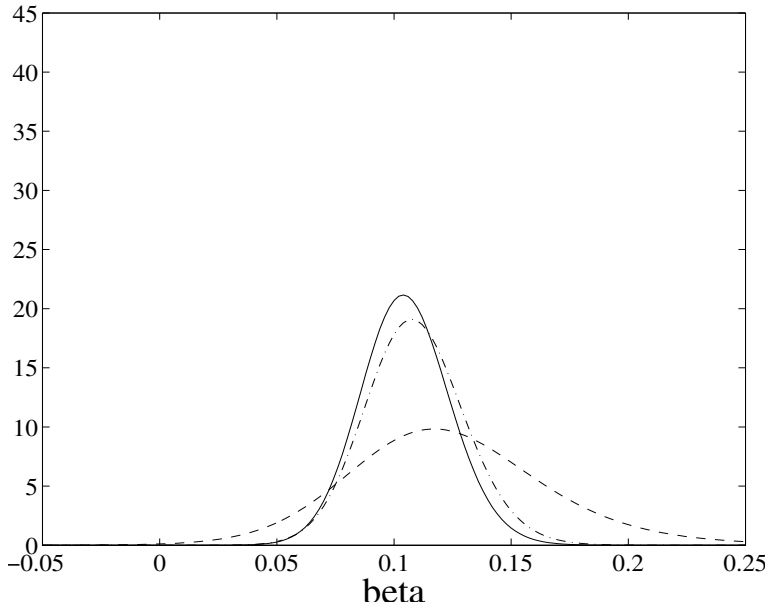

Figure 5.8: Marginal posterior of return on education β under Jeffreys prior for data of Kentucky, Arkansas and Tennessee: original model (solid), model with first quarter dummy in wage equation (dash-dot), model with three quarter dummies in wage equation (dashed).

Finally, Bound, Jaeger and Baker (1995) note that random instruments yield results similar to those for real data for four model specifications. For each specification the (mean) asymptotic standard error (over 500 simulations) of the 2SLS estimator for β is only somewhat larger for random data than for real data: 2.3, 1.3, 1.7 and 1.4 times larger. For our specification the asymptotic standard error of the 2SLS estimator for β is also only 1.5 times larger for random instruments than for real instruments (for the whole data set of the US). However, the 95% posterior interval under the Jeffreys prior is 3.0 times wider for random instruments than for real instruments, so the use of the Jeffreys prior shows a clear difference between results for random and real data. This reflects the relative insensitivity of Bayesian analysis under the Jeffreys prior to the addition of irrelevant instruments as compared to the flat prior (and the 2SLS estimator), as mentioned by Kleibergen and Zivot (2003).

5.6 Conclusions

We have shown results of two classical methods, the two-stage least squares (2SLS) and limited information maximum likelihood (LIML) estimators, and two Bayesian approaches, using the flat prior of Drèze (1976) and the Jeffreys prior, for an IV regression model of Angrist and Krueger (1991) for the return on education. It is shown that for these four methods the results for the US crucially depend on the results for the Census region South. A possible explanation for this is that the average education spell for men born in 1930-1939 is lower in the South, implying a larger influence of compulsory schooling laws as these do not concern education above a certain number of years, and hence a stronger effect of quarter of birth on education. A further division shows that results for the South substantially depend on three states: Kentucky, Tennessee and Arkansas. This suggests that the average level of education is not the only factor influencing the strength of the instruments, as men born in Alabama, Mississippi, Virginia and West Virginia have on average completed less years of education than those born in Tennessee: there are also other factors playing a role, which may include the power of government agencies enforcing schooling laws and the exemptions from these schooling laws, which vary across states.

If the effect of the return on education differs between the four Census regions, which may not a priori be ruled out given the large economic differences between these regions, inference using data of the US is not representative for the average returns on education across the US. Therefore one should be careful when drawing such conclusions.

We have further shown that quarter of birth is a stronger instrument for education for people with at most 8 or at least 14 years of education than for people with 9-13 years of education, which suggests that quarter of birth does not only affect the number of completed years of schooling for those who leave school as soon as it is allowed, as these are (mostly) contained in the group with 9-13 years of education. This suggests that the strength of the quarter of birth instruments is not so much caused by the asymmetry between school entry requirements and compulsory schooling laws keeping students at school until they reach a certain age; the value of the quarter of birth instruments seems to stem to a larger extent from the school entry requirements in combination with the dependence of the 'hazard rate' of leaving school on age (measured in quarters). Therefore, if one intends to increase the understanding of the working of the quarter-of-birth instruments, it is probably a better idea to pay more attention to school entry require-

ments and/or compulsory schooling laws for children of age 5-7 than to concentrate on the differences between compulsory schooling laws for students of age 16-18.

Finally, Bound, Jaeger and Baker (1995) have concluded that the interaction between compulsory school attendance laws and quarter of birth does not give much usable information concerning the causal effect of education on wages for two main reasons. First, the weakness of the instruments may lead to large inconsistencies in the 2SLS estimator even if there is only a weak relationship between the instruments and the error in the structural equation; Bound, Jaeger and Baker (1995) mention evidence casting doubt on the assumption that no such correlation is present. Moreover, Bound, Jaeger and Baker (1995) even report that differences in family income at time of birth would seem to account for virtually all of the association between quarter of birth and wages: they argue that the difference in income between those born in the first quarter and those born during the rest of the year can almost completely be explained by differences in family income at time of birth and an intergenerational correlation. Second, the 2SLS estimates reported by Angrist and Krueger (1991) may suffer from substantial finite sample biases because of the weakness of the instruments (despite the large sample size). However, we can use a Bayesian approach under the Jeffreys prior or the LIML estimator, which is approximately median unbiased in this case, instead of the 2SLS estimator. Furthermore, we may still obtain a rather tight posterior for β if we allow for a direct effect of birth during the first quarter on income. It should be noted that including quarter-of-birth dummies in the wage equation may result in (much) wider posterior intervals, and that if a direct effect of quarter of birth on income exists, this may not be constant across states/years; in that case more terms should be added to the wage equation. So, an important question remains whether the inclusion of such terms in the wage equation is necessary and if so, how these should be specified. This is left as a topic for further research. Still, it should at least be noted that if there exists a direct effect of quarter of birth on income, it is not likely that the factors causing this effect differ between states/years in the same way as compulsory schooling laws and the degree to which these are enforced. So, even if there exists a direct effect of quarter of birth on income which varies across states/years, the difference between these effects and the effect of compulsory schooling laws may be exploited, so that the resulting model may still give usable information on the causal effect of education on income in (regions of) the US.

So, it seems that the conclusion of Bound, Jaeger and Baker (1995), that the interaction between compulsory school attendance laws and quarter of birth does not give much usable information concerning the causal effect of education on earnings, may have been

too strong, as the model of Angrist and Krueger (1991) (or a slightly modified version) may give usable information on the causal effect of education on income in (regions of) the US.

We end this chapter mentioning two topics for further research. First, an obvious question is whether the results reported in this chapter can also be found for other model specifications considered by Angrist and Krueger (1991). Second, an interesting idea is to apply the approaches used in this chapter under the assumption that the disturbances obey a different distribution than the normal, and thus investigate whether the results in this chapter are robust with respect to this distributional assumption.

Chapter 6

Summary and further research

In this thesis a class of neural network sampling methods is introduced and analyzed, and several issues concerning instrumental variables regression models are investigated. In this chapter the main findings of this thesis are summarized and some topics for further research are mentioned.

In Chapter 2 a class of neural network sampling algorithms is introduced that can be useful when one needs to evaluate high-dimensional integrals in cases of highly non-elliptical target (posterior) distributions. In these algorithms a neural network function is used as an importance or candidate density in the importance sampling (IS) or (independence chain) Metropolis-Hastings (MH) algorithm. Neural networks are natural importance or candidate densities, as they have a universal approximation property and are easy to sample from. We have shown how to sample from three types of neural networks. One can sample directly from a certain 3-layer network. Using a 4-layer network one can, depending on the specification of the network, either use a Gibbs sampling approach or sample directly from a mixture of distributions. A key step in the proposed class of methods is the construction of a neural network that approximates the target density accurately. The methods have been tested on an illustrative example; the 4-layer network specified as the mixture of t distributions performed the best among the proposed sampling procedures. In another experiment concerning a bimodal posterior distribution in an IV regression for a simulated data set, the approach using a mixture of t distributions provided (in the same computing time) more accurate results than IS with a unimodal importance density or a random walk Metropolis-Hastings algorithm. The Gibbs sampler failed in this example, as it got stuck in one of the modes. These results indicate the feasibility and the possible usefulness of the neural network approach.

The proposed techniques can be extended in the following ways. First, one may use these results in model selection and model averaging and investigate the effect of using accurate non-elliptical credible sets instead of naive or asymptotic sets.

Second, one may consider other ways of specifying and estimating neural networks. One may also, as a first step, transform the posterior density function to a more regular shape. This line of research is recently pursued by *e.g.* Bauwens et al. (2004) in a class of adaptive direction sampling methods using radial-basis functions (ARDS). A combination of ADS and neural network sampling may be of interest. In practice, one also encounters cases where only part of the posterior density is ill-behaved. Then one may combine the neural network approach for the 'difficult part' with a Gibbs sampling approach for the regular part of the model. Another area of further research is to consider different flexible candidate density functions involving Hermite polynomials, see *e.g.* Gallant and Tauchen (1993) and the references cited there. Also, more sophisticated Monte Carlo methods like bridge sampling, see *e.g.* Meng and Wong (1996) and Frühwirth-Schnatter (2004), may be explored in combination with neural networks.

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

Fourth, the neural network approximations proposed in this thesis may be useful for modelling such processes as volatility in financial series, see *e.g.* Donaldson and Kamstra (1997), and for evaluating option prices, see Hutchinson, Lo and Poggio (1994).

Chapter 3 discussed some large improvements in the AdMit method, in which an Adaptive Mixture of t distributions is used as a candidate distribution in the IS or MH method; these improvements make the method faster (about three times as fast in an example of a 3-dimensional bimodal target distribution) and more reliable (in the sense of a quicker detection of distant modes) as compared to the method proposed in Chapter 2. The improved AdMit methods are applied to a 4-dimensional posterior distribution in a mixture model for US real GNP growth rates. The AdMit methods outperform two Gibbs sampling approaches, Gibbs sampling with data augmentation and the griddy Gibbs sampler; in this case the Gibbs sequences did not get stuck in one of two modes – in fact the joint posterior density is unimodal in this example – but the high serial correlation in the Gibbs sequences caused the Gibbs samplers to yield estimators of posterior moments

with larger standard deviations than those resulting from the neural network methods (in the same computing time). Finally, it is illustrated that neural network sampling methods can especially be useful if one desires estimators of posterior characteristics with high precision.

The proposed techniques can be extended in the following ways. A straightforward alternative is to use a neural network function as a candidate density in rejection sampling instead of importance sampling or the Metropolis-Hastings algorithm. Another extension that is more difficult to implement, but much more interesting for practical purposes is to build a neural network method within a Gibbs sampling procedure (or a 'MH within Gibbs' algorithm). If it is hard to draw from one of the conditional distributions, say the conditional distribution of θ^a given θ^b with $\theta = (\theta^a, \theta^b)$ where θ^a and θ^b both consist of multiple elements, two options are to use a 'MH within Gibbs' step or to use several steps of the griddy Gibbs sampler. For a 'MH within Gibbs' step a candidate density is required. An option is to approximate the conditional target density of θ^a given θ^b with a mixture of Student-t densities. However, the disadvantage is that in each iteration (for each different value of θ^b) a new approximation has to be constructed, which can result in a very time consuming algorithm. In order to keep the computing time for obtaining approximations to conditional target densities relatively small, one can store both the θ^b's and the approximations to the conditional densities of θ^a given θ^b. In each iteration one can use as an initial point the (mixture of t) approximation for the value of θ^b that is closest to the current value of θ^b (taking into account the scales and correlations of the elements of θ^b) among the set of previous θ^b's in the Markov chain. After that, one may add one or more components to the candidate mixture and drop (almost) useless Student-t components in order to prevent ending up with mixtures of huge numbers of components. Nevertheless, the resulting algorithm will still be rather slow. However, a 'MH withing Gibbs' step with a poor candidate distribution may result in a very low acceptance probability, resulting in very slow convergence of the estimators, or even an unreliable algorithm in which certain regions of the domain of θ that contain substantial probability mass may be 'missed'. And the use of several griddy Gibbs steps also yields a slow algorithm, in which the division of the sampling of θ^a into individual steps for sampling the elements of θ^a may seriously increase the serial correlation in the Gibbs sequence. Therefore, the combination of neural network sampling methods and the Gibbs sampler (or the 'MH within Gibbs' algorithm) is an interesting topic for further research.

In Chapter 4 it has been considered how shapes of posteriors in the instrumental variables (IV) regression model under the flat or Jeffreys prior depend on the level of endogeneity and instrument strength. Further, it is considered how the Jeffreys prior 'remedies' two of the peculiar properties of the posterior under the flat prior, the asymptote of the marginal posterior of Π at $\Pi = 0$ and the dependence of the tail behavior of the marginal posterior of β on the number of instruments in the sense that its tails become thinner when (possibly irrelevant) instruments are added to the model. For the case of one explanatory endogenous variable an explicit formula exists for the marginal posterior of β under the Jeffreys prior, see *e.g.* Kleibergen and Zivot (2003). If there are m explanatory endogenous variables with $m > 1$, then sampling methods are required. Kleibergen and van Dijk (1998) and Kleibergen and Paap (2002) have derived importance sampling and Metropolis-Hastings algorithms that are specifically designed for reduced rank regression models such as the IV regression model. However, in the case of many instruments these may require the evaluation of a determinant of a huge Jacobian matrix, which may be numerically cumbersome. If these sampling methods are not applicable in certain cases, the possibility of (highly) non-elliptical shapes of the posterior distributions under the Jeffreys prior implies that neural network sampling methods may be useful tools in a Bayesian analysis of an IV model under the Jeffreys prior, possibly after some parameter transformations. This is left as a topic for further research.

Finally, the hierarchical prior of Chamberlain and Imbens (1996) is briefly discussed, which can also be considered as a 'regularization prior' that 'cures' strange posterior properties occurring under the flat prior. Unlike the approach under the flat prior, the approach of Chamberlain and Imbens (1996) is also capable of only resulting in tight HPD regions for β in the case of data sets that contain information on β, just like the approach using the Jeffreys prior. The hierarchically based prior in the approach of Chamberlain and Imbens (1996) requires the 'tuning' of some prior variance (or covariance matrix), which is a major disadvantage as compared to the approach under the Jeffreys prior. It should be noted that the approach of Chamberlain and Imbens (1996) has the advantage that the hierarchical prior is not necessarily data dependent, while the Jeffreys prior generally is, and that a straightforward Gibbs sampler can be used to sample from the corresponding posterior. However, the disadvantage of the 'tuning' of a prior variance, and the sensitivity of posterior results to the choice of this prior variance, clearly suggest that the use of the Jeffreys prior is preferable in most situations.

Finally, in Chapter 5 the results are shown of two classical methods, the two-stage least squares (2SLS) and limited information maximum likelihood (LIML) estimators, and two Bayesian approaches, using the flat prior of Drèze (1976) and the Jeffreys prior, for an IV regression model of Angrist and Krueger (1991) for the return on education, in which quarter of birth is used to construct instrumental variables. It is shown that for these four methods the results for the US (for men born in 1930-1939) crucially depend on the results for the Census region South. A possible explanation for this is that the average education spell for men born in 1930-1939 is lower in the South, implying a larger influence of compulsory schooling laws as these do not concern education above a certain number of years, and hence a stronger effect of quarter of birth on education. A further division shows that results for the South substantially depend on three states: Kentucky, Tennessee and Arkansas. This suggests that the average level of education is not the only factor influencing the strength of the instruments, as men born in Alabama, Mississippi, Virginia and West Virginia have on average completed less years of education than those born in Tennessee: there are also other factors playing a role, which may include the power of government agencies enforcing schooling laws and the exemptions from these schooling laws, which vary across states.

If the effect of the return on education differs between the four Census regions, which may not a priori be ruled out given the large economic differences between these regions, inference using data of the US is not representative for the average returns on education across the US. Therefore one should be careful when drawing such conclusions.

We have further shown that quarter of birth is a stronger instrument for education for people with at most 8 or at least 14 years of education than for people with 9-13 years of education, which suggests that quarter of birth does not only affect the number of completed years of schooling for those who leave school as soon as it is allowed, as these are (mostly) contained in the group with 9-13 years of education. This suggests that the strength of the quarter of birth instruments is not so much caused by the asymmetry between school entry requirements and compulsory schooling laws keeping students at school until they reach a certain age, which is the reason suggested by Angrist and Krueger (1991). The value of the quarter of birth instruments seems to stem to a larger extent from the school entry requirements in combination with the dependence of the 'hazard rate' of leaving school on age (measured in quarters). Therefore, if one intends to increase the understanding of the working of the quarter-of-birth instruments, it is probably a better idea to pay more attention to school entry requirements and/or compulsory schooling

laws for children of age 5-7 than to concentrate on the differences between compulsory schooling laws for students of age 16-18.

Finally, the criticism of Bound, Jaeger and Baker (1995) is discussed. Bound, Jaeger and Baker (1995) have concluded that the interaction between compulsory school attendance laws and quarter of birth does not give much usable information concerning the causal effect of education on wages for two main reasons. First, the weakness of the instruments may lead to large inconsistencies in the 2SLS estimator even if there is only a weak relationship between the instruments and the error in the structural equation; Bound, Jaeger and Baker (1995) mention evidence casting doubt on the assumption that no such correlation is present. Moreover, Bound, Jaeger and Baker (1995) even report that differences in family income at time of birth would seem to account for virtually all of the association between quarter of birth and wages: they argue that the difference in income between those born in the first quarter and those born during the rest of the year can almost completely be explained by differences in family income at time of birth and an intergenerational correlation. Second, the 2SLS estimates reported by Angrist and Krueger (1991) may suffer from substantial finite sample biases because of the weakness of the instruments (despite the large sample size). However, we can use a Bayesian approach under the Jeffreys prior or the LIML estimator, which is approximately median unbiased in this case, instead of the 2SLS estimator. Furthermore, we may still obtain a rather tight posterior for β if we allow for a direct effect of birth during the first quarter on income. It should be noted that including quarter-of-birth dummies in the wage equation may result in (much) wider posterior intervals, and that if a direct effect of quarter of birth on income exists, this may not be constant across states/years; in that case more terms should be added to the wage equation. So, an important question remains whether the inclusion of such terms in the wage equation is necessary and if so, how these should be specified. This is left as a topic for further research. Still, it should at least be noted that if there exists a direct effect of quarter of birth on income, it is not likely that the factors causing this effect differ between states/years in the same way as compulsory schooling laws and the degree to which these are enforced. So, even if there exists a direct effect of quarter of birth on income which varies across states/years, the difference between these effects and the effect of compulsory schooling laws may be exploited, so that the resulting model may still give usable information on the causal effect of education on income in (regions of) the US.

So, it seems that the conclusion of Bound, Jaeger and Baker (1995), that the interaction between compulsory school attendance laws and quarter of birth does not give much

usable information concerning the causal effect of education on earnings, may have been too strong, as the model of Angrist and Krueger (1991) (or a slightly modified version) may give usable information on the causal effect of education on income in (regions of) the US.

An obvious question is whether the results reported in this chapter can also be found for other model specifications considered by Angrist and Krueger (1991). Another interesting idea is to apply the approaches used in Chapter 5 under the assumption that the disturbances obey a different distribution than the normal, and thus investigate whether the results in this chapter are robust with respect to this distributional assumption. These issues are left as two topics for further research.

Samenvatting
(Summary in Dutch)

Het onderzoek in dit proefschrift betreft ruw gesteld drie onderwerpen: neurale netwerken, simulatie methoden en regressie modellen met instrumentele variabelen (IV). Er bestaat een verband tussen deze onderwerpen, dat hieronder uitgelegd zal worden. Eerst dient opgemerkt te worden dat, hoewel in hoofdstuk 5 de resultaten van de Bayesiaanse benadering vergeleken worden met die van klassieke methoden, in dit proefschrift de nadruk ligt op de Bayesiaanse benadering van econometrie/statistiek. In de Bayesiaanse benadering is het in het algemeen noodzakelijk om integralen van de posterior kansdichtheid over de parameters van het model te berekenen. Als deze integralen niet analytisch geëvalueerd kunnen worden, en de dimensie van het integratie probleem groter is dan 3 of 4, zodat de mogelijkheid van toepassing van deterministische integratie methoden wegvalt, dan heeft men Monte Carlo integratie methoden nodig.

Twee bekende Monte Carlo integratie methoden zijn importance sampling (IS) en het Metropolis-Hastings (MH) algoritme. In deze methoden worden trekkingen gedaan uit een zekere kandidaat verdeling, en het verschil tussen deze kandidaat verdeling en de posterior kansverdeling waarin men geïnteresseerd is wordt 'gecorrigeerd' door trekkingen te wegen (in IS), of door een Markov keten te construeren waarbij bepaalde kandidaat trekkingen verworpen worden en andere trekkingen een aantal keer geaccepteerd worden (in het MH algoritme). De kansen op verwerping/acceptatie worden hierbij zo gekozen dat de kansverdeling van de elementen in de Markov keten convergeert naar de posterior kansverdeling.

Een gebruikelijke keuze voor een kandidaat verdeling is een normale of Student-t verdeling. Echter, als het verschil tussen de kandidaat en posterior kansdichtheid groot is, bijvoorbeeld als de schaal en/of de locatie van de modus veel verschillen, dan kunnen de IS en MH methoden langzaam zijn of onbetrouwbare resultaten geven: er kunnen veel trekkingen nodig zijn om tot convergentie te komen van de schattingen van de eigen-

schappen van de posterior kansverdeling waarin men geïnteresseerd is, of in het geval van multi-modaliteit kan het voorkomen dat er een modus 'gemist' wordt. Als de posterior kansverdeling (sterk) niet-elliptisch is, bijvoorbeeld in het geval van multi-modaliteit, dan heeft men dus in sommige gevallen een andere kandidaat verdeling nodig, bij voorkeur een kandidaat verdeling die (bij benadering) dezelfde afwijkende eigenschappen als de posterior kansverdeling heeft.

Een typische reden voor sterk niet-elliptische posterior kansverdelingen is de aanwezigheid van locale non-identificatie, het verschijnsel dat voor bepaalde waarden van sommige parameters andere parameters niet geïdentificeerd zijn. In dit proefschrift worden twee modellen beschouwd waarin locale non-identificatie een rol speelt, het IV regressie model en een model met 2 regimes.

In het geval van een sterk niet-elliptische posterior kansverdeling kan een verstandig gekozen niet-elliptische kandidaat verdeling een grote verbetering betekenen ten opzichte van een normale of Student-t kandidaat verdeling. In zulke gevallen kan het gebruik van neurale netwerken als kandidaat dichtheid nuttig zijn.

De simulatie methoden op basis van neural netwerken die in dit proefschrift geïntroduceerd worden bestaan uit twee stappen. Eerst wordt een neuraal netwerk geconstrueerd dat de posterior verdeling benadert, waarna dit neurale netwerk gebruikt wordt als kandidaat kansdichtheid in de IS of MH methode. Dit betekent dat we neurale netwerken moeten gebruiken waaruit, wanneer deze netwerken als kansdichtheid beschouwd worden, eenvoudig trekkingen gedaan kunnen worden. De klasse van neurale netwerken die wij beschouwen bevat convexe combinaties van Student-t verdelingen, waaruit eenvoudig trekkingen gedaan kunnen worden. Uit deze convexe combinaties van Student-t verdelingen kan niet alleen eenvoudig getrokken worden, deze klasse van verdelingen is ook zeer flexibel in de zin dat een breed spectrum van (posterior) kansdichtheden door deze verdelingen benaderd kan worden.

Sterk niet-elliptische posterior kansverdelingen kunnen voorkomen in IV regressie modellen, wat het bovengenoemde verband tussen simulatie methoden op basis van neurale netwerken en IV regressie modellen verklaart. De vorm van de posterior kansverdeling in het IV regressie model onder een platte prior of Jeffreys prior hangt af van de verklarende kracht van de instrumentele variabelen, en van de mate van endogeniteit in het model. Sterk niet-elliptische posterior kansverdelingen worden vooral verkregen in het geval van zwakke instrumentele variabelen.

In dit proefschrift is verder uitgebreid aandacht besteed aan het bekende IV regressie model van Angrist en Krueger (1991) voor het effect van (het aantal jaren) onderwijs op het inkomen van individuen. Kennis over dit effect is onder andere relevant voor overheden die besluiten nemen over de wetten betreffende de leerplicht. Vanwege de endogeniteit van onderwijs en inkomen gebruiken Angrist en Krueger (1991) instrumentele variabelen die geconstrueerd zijn op basis van het kwartaal waarin iemand geboren is. Het is moelijk om instrumenten te vinden die gecorreleerd zijn met het aantal jaren onderwijs dat iemand voltooid heeft, maar ongecorreleerd met niet waargenomen 'capaciteiten' die zowel invloed hebben op onderwijs als op inkomen. Het schatten van het daadwerkelijke effect van onderwijs op inkomen is daarom een niet-triviale zaak. De instrumenten, die gebaseerd zijn op het kwartaal waarin iemand geboren is, maken gebruik van het verschil in de gemiddelde hoeveelheid onderwijs tussen mensen geboren in verschillende kwartalen. Angrist en Krueger (1991) redeneren als volgt. De meeste school districten in de Verenigde Staten van Amerika eisen dat kinderen 6 jaar zijn op 1 januari van het jaar waarin ze voor het eerst naar school gaan, terwijl kinderen op school moeten blijven tot hun zestiende, zeventiende of achttiende verjaardag. Deze asymmetrie tussen toelatingseisen van scholen en leerplicht wetten voor tieners zorgt ervoor dat scholieren die in bepaalde maanden geboren zijn, langer verplicht zijn naar school te gaan dan scholieren die in andere maanden geboren zijn: scholieren die eerder in het jaar geboren zijn, gaan op een hogere leeftijd voor het eerst naar school en bereiken de leeftijd waarop ze van school mogen gaan na minder onderwijs. Daarom geldt voor scholieren die van school gaan zodra het wettelijk toegestaan is, dat degenen die in het eerste kwartaal geboren zijn, gemiddeld genomen drie kwartalen minder onderwijs hebben gehad dan degenen die in het vierde kwartaal geboren zijn.

In dit proefschrift worden de resultaten beschouwd van een Bayesiaanse benadering met een platte of Jeffreys prior, en van klassieke benaderingen met de twee-staps kleinste kwadraten schatter of de maximale aannemelijkheids schatter. Er wordt getoond dat de resultaten van deze vier benaderingen voor data van de V.S. gedomineerd worden door de data van de zuidelijke staten. Als het effect van onderwijs op inkomen verschilt tussen de regio's van de V.S., wat mogelijk niet a priori uitgesloten mag worden gezien de grote economische verschillen tussen deze regio's, dan is het geschatte effect van onderwijs op inkomen niet representatief voor het gemiddelde effect van onderwijs op inkomen over de V.S. Daarom moet men voorzichtig zijn bij het trekken van dergelijke conclusies.

Bound, Jaeger en Baker (1995) hebben geconcludeerd dat de interactie tussen de leerplicht wetten en kwartalen waarin personen geboren zijn, de basis van de modellen van

Angrist en Krueger (1991), niet veel bruikbare informatie over het effect van onderwijs op inkomen geeft. Als redenen hiervoor worden aangevoerd dat in het geval van dergelijk zwakke instrumenten de twee-staps kleinste kwadraten schatter onzuiver is, en dat in dit geval een kleine correlatie tussen de instrumenten en de storingsterm in de structurele vergelijking voor grote inconsistenties zorgt. Daarbij worden redenen gegeven om te twijfelen aan de afwezigheid van een dergelijke correlatie; vooral wordt in twijfel getrokken dat het verschil in inkomen tussen personen die in het eerste kwartaal geboren zijn en degenen die in de andere drie kwartalen geboren zijn, uitsluitend door verschillen in onderwijs veroorzaakt is. Echter, we kunnen in plaats van de twee-staps kleinste kwadraten schatter gebruik maken van een Bayesiaanse methode op basis van de Jeffreys prior of de maximale aannemelijkheids schatter die in dit geval (bij benadering) mediaan zuiver is. Bovendien kunnen we ook een plausibele posterior kansverdeling met een tamelijke kleine spreiding verkrijgen voor het effect van onderwijs op inkomen, als we een direct effect van geboorte in het eerste kwartaal op het inkomen toelaten. Dit suggereert dat de conclusie van Bound, Jaeger en Baker (1995) te sterk is, omdat het model van Angrist en Krueger (1991) (of een licht aangepaste versie) bruikbare informatie kan geven over het causale effect van onderwijs op inkomen in (regio's van) de V.S.

Bibliography

Anderson, T.W., Rubin, H., 1949. Estimators of the parameters of a single equation in a complete set of stochastic equations. The Annals of Mathematical Statistics 21, 570–582.

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

Angrist, J.D., Krueger, A.B., 1992. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. Journal of the American Statistical Association 87, 328–336.

Basmann, R.L., 1957. A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica 25, 77–83.

Basmann, R.L., 1963. Remarks concerning the application of exact finite sample distribution functions of GCL estimators in econometric statistical inference. Journal of the American Statistical Association 58, 943–976.

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

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.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press.

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.

Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association 90, 443–450.

Casella, G., George, E.I., 1992. Explaining the Gibbs sampler. The American Statistician 46, 167–174.

Chamberlain, G., Imbens, G.W., 1996. Hierarchical Bayes models with many instrumental variables. NBER Technical Working Paper 204.

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

Cheney, W., Kincaid, D., 1994. Numerical Mathematics and Computing, third edition, Brooks-Cole.

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.

Crespin, D., 1995. Neural network formalism. Manuscript, available at `http://euler.ciens.ucv.ve/~dcrespin/Pub/formal.html`.

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

Defense Advanced Research Projects Agency (DARPA), 1988. DARPA Neural Network Study: October 1987-February 1988. AFCEA International Press, Fairfax, Va., USA.

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

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

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

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

Fiesler, E., 1994. Neural network classification and formalization. Computer Standards and Interfaces 16, 231–239.

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

Gallant, A.R., Tauchen, G., 1993. A nonparametric approach to nonlinear time series analysis: estimation and simulation. In: Brillinger, D., Caines, P., Geweke, J., Parzen, E., Rosenblatt, M., Taqqu, M.S. (Eds.), New Directions in Time Series Analysis Part II, Springer-Verlag, New York.

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.

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

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

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

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

Hausman, J.A., 1983. Specification and estimation of simultaneous equations systems. In: Griliches, Z., Intrilligator, M.D., editors, Handbook of Econometrics, volume 1. Elsevier Science, Amsterdam.

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.

Hood, W.C., Koopmans, T.C., 1953. Studies in Econometric Method, volume 14 of Cowles Foundation Monograph. Wiley, New York.

Hoogerheide, L.F., Van Dijk, H.K., 2001. Comparison of the Anderson-Rubin test for overidentification and the Johansen test for cointegration. Econometric Institute report EI 2001-04, Erasmus University Rotterdam.

Hoogerheide, L.F., Kaashoek, J.F., Van Dijk, H.K., 2003a. Functional approximations to posterior densities: a neural network approach to efficient sampling. Econometric Institute report EI 2002-48, Erasmus University Rotterdam.

Hoogerheide, L.F., Kaashoek, J.F, Van Dijk, H.K., 2003b. Neural network approximations to posterior densities: an analytical approach. 2003 Proceedings: Bayesian Statistical Science, American Statistical Association, pp. 1857–1862.

Hoogerheide, L.F., Kaashoek, J.F., Van Dijk, H.K., 2004. Neural network based approximations to posterior densities: a class of flexible sampling methods with applications to reduced rank models. Econometric Institute report EI 2004-19, Erasmus University Rotterdam.

Hoogerheide, L.F, Kaashoek, J.F., Van Dijk, H.K., 2006. 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, forthcoming.

Hoogerheide, L.F, Kleibergen, F.R., Van Dijk, H.K., 2006. Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. Journal of Econometrics, forthcoming.

Hoogerheide, L.F, Van Dijk, H.K., 2006a. Searching for TOPS: The Optimal Posterior Simulator. Econometric Institute report EI 2006-15, Erasmus University Rotterdam.

Hoogerheide, L.F., and Van Dijk, H.K., 2006b. A reconsideration of the Angrist-Krueger analysis on returns to education. Manuscript in preparation.

Hoogerheide, L.F, Van Dijk, H.K., Van Oest, R.D., 2006. Simulation methods for Bayesian Econometric Inference, Chapter in Handbook of Computational Economics and Statistics. Elsevier Publishing, in preparation.

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

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

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

Kleibergen, F.R., 2000. Exact test statistics and distributions of maximum likelihood estimators that result from orthogonal parameters with applications to the instrumental variables regression model. Discussion paper TI 2000-039/4, Tinbergen Institute, Rotterdam.

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

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

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

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

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

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

Kolmogorov, A.N., 1957. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. American Mathematical Monthly Translation 28, 55–59. (Russian original in Doklady Akademii Nauk SSSR, 144, 953–956)

Lancaster, T., 2004. An Introduction to Modern Bayesian Econometrics. Blackwell Publishing, Oxford.

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

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

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

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

Paap, R., Van Dijk, H.K., 2003. Bayes estimates of Markov trends in possibly cointegrated series: an application to US consumption and income. Journal of Business & Economic Statistics 21, 547–563.

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

Phillips, P.C.B., 1983. Exact small sample theory in the simultaneous equations model. In Griliches, Z., Intrilligator, M.D., editors, Handbook of Econometrics, Vol.1. North-Holland Publishing Co., Amsterdam.

Ritter, C., Tanner, M.A., 1992. Facilitating the Gibbs sampler: the Gibbs stopper and the Griddy-Gibbs sampler. Journal of the American Statistical Association 87, 861–868.

Rosenblatt, F., 1962. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan, Washington DC.

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

Solon, G., 1992. Intergenerational income mobility in the United States. American Economic Review 82, 393–408.

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

Stergiou, C., Siganos, D., 1996. Neural networks. Manuscript, available at `http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html`.

Stinchcombe, M., White, H., 1989. Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In: Proceedings of the International Joint Conference on Neural Networks, Washington DC, IEEE Press, New York.

Stinchcombe, M., White, H., 1990. Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights. In: Proceedings of the International Joint Conference on Neural Networks, IEEE Press, Piscataway, New Jersey.

Stoer, J., Bulirsch, R., 1993. Introduction to Numerical Analysis, second edition, Springer-Verlag.

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

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

Theil, H., 1953. Estimation and Simultaneous Correlation in Complete Equation Systems. Mimeographed Memorandum of the Central Planning Bureau, The Hague.

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

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

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

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

Van Oest, R.D., 2005. Essays on Quantitative Marketing Models and Monte Carlo Integration Methods. Ph.D. thesis, Tinbergen Institute, Rotterdam.

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.

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

Zimmerman, D.J., 1992. Regression toward mediocrity in economic stature. American Economic Review 82, 409–429.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Rotterdam and Amsterdam. The following books recently appeared in the Tinbergen Institute Research Series:

328. W. VAN WINDEN, *Essays on urban ICT policies.*
329. G.J. KULA, *Optimal retirement decision.*
330. R.J. IMESON, *Economic analysis and modeling of fisheries management in complex marine ecosystems.*
331. M. DEKKER, *Risk, resettlement and relations: Social security in rural Zimbabwe.*
332. A. MULATU, *Relative stringency of environmental regulation and international competitiveness.*
333. C.M. VAN VEELEN, *Survival of the fair: Modelling the evolution of altruism, fairness and morality.*
334. R. PHISALAPHONG, *The impact of economic integration programs on inward foreign direct investment.*
335. A.H. NÖTEBERG, *The medium matters: The impact of electronic communication media and evidence strength on belief revision during auditor-client inquiry.*
336. M. MASTROGIACOMO, *Retirement, expectations and realizations. Essays on the Netherlands and Italy.*
337. E. KENJOH, *Balancing work and family life in Japan and four European countries: Econometric analyses on mothers' employment and timing of maternity.*
338. A.H. BRUMMANS, *Adoption and diffusion of EDI in multilateral networks of organizations.*
339. K. STAAL, *Voting, public goods and violence.*
340. R.H.J. MOSCH, *The economic effects of trust. Theory and empirical evidence.*
341. F. ESCHENBACH, *The impact of banks and asset markets on economic growth and fiscal stability.*
342. D. LI, *On extreme value approximation to tails of distribution functions.*
343. S. VAN DER HOOG, *Micro-economic disequilibrium dynamics.*
344. B. BRYS, *Tax-arbitrage in the Netherlands evaluation of the capital income tax reform of January 1, 2001.*

345. V. PRUZHANSKY, *Topics in game theory.*

346. P.D.M.L. CARDOSO, *The future of old-age pensions: Its implosion and explosion.*

347. C.J.H. BOSSINK, *To go or not to go? International relocation willingness of dual-career couples.*

348. R.D. VAN OEST, *Essays on quantitative marketing models and Monte Carlo integration methods.*

349. H.A. ROJAS-ROMAGOSA, *Essays on trade and equity.*

350. A.J. VAN STEL, *Entrepreneurship and economic growth: Some empirical studies.*

351. R. ANGLINGKUSUMO, *Preparatory studies for inflation targeting in post crisis Indonesia.*

352. A. GALEOTTI, *On social and economic networks.*

353. Y.C. CHEUNG, *Essays on European bond markets.*

354. A. ULE, *Exclusion and cooperation in networks.*

355. I.S. SCHINDELE, *Three essays on venture capital contracting.*

356. C.M. VAN DER HEIDE, *An economic analysis of nature policy.*

357. Y. HU, *Essays on labour economics: Empirical studies on wage differentials across categories of working hours, employment contracts, gender and cohorts.*

358. S. LONGHI, *Open regional labour markets and socio-economic developments: Studies on adjustment and spatial interaction.*

359. K.J. BENIERS, *The quality of political decision making: Information and motivation.*

360. R.J.A. LAEVEN, *Essays on risk measures and stochastic dependence: With applications to insurance and finance.*

361. N. VAN HOREN, *Economic effects of financial integration for developing countries.*

362. J.J.A. KAMPHORST, *Networks and learning.*

363. E. PORRAS MUSALEM, *Inventory theory in practice: Joint replenishments and spare parts control.*

364. M. ABREU, *Spatial determinants of economic growth and technology diffusion.*

365. S.M. BAJDECHI-RAITA, *The risk of investment in human capital.*

366. A.P.C. VAN DER PLOEG, *Stochastic volatility and the pricing of financial derivatives.*

367. R. VAN DER KRUK, *Hedonic valuation of Dutch Wetlands.*

368. P. WRASAI, *Agency problems in political decision making.*

369. B.K. BIERUT, *Essays on the making and implementation of monetary policy decisions.*

370. E. REUBEN, *Fairness in the lab: The effects of norm enforcement in economic decisions.*

371. G.J.M. LINDERS, *Intangible barriers to trade: The impact of institutions, culture, and distance on patterns of trade.*

372. A. HOPFENSITZ, *The role of affect in reciprocity and risk taking: Experimental studies of economic behavior.*

373. R.A. SPARROW, *Health, education and economic crisis: Protecting the poor in Indonesia.*

374. M.J. KOETSE, *Determinants of investment behaviour: Methods and applications of meta-analysis.*

375. G. MÜLLER, *On the role of personality traits and social skills in adult economic attainment.*

376. E.H.B. FEIJEN, *The influence of powerful firms on financial markets.*

377. J.W. GROSSER, *Voting in the laboratory.*

378. M.R.E. BRONS, *Meta-analytical studies in transport economics: Methodology and applications.*