EUI Working Papers

ECO 2010/27
DEPARTMENT OF ECONOMICS

ESTIMATION METHODS COMPARISON OF SVAR MODEL WITH THE MIXTURE OF TWO NORMAL DISTRIBUTIONS – MONTE CARLO ANALYSIS

Katarzyna Maciejowska

EUROPEAN UNIVERSITY INSTITUTE, FLORENCE
DEPARTMENT OF ECONOMICS

Estimation Methods Comparison of SVAR Model with the Mixture of Two Normal Distributions – Monte Carlo Analysis

KATARZYNA MACIEJOWSKA

EUI Working Paper **ECO** 2010/27

ISSN 1725-6704

Printed in Italy
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy
www.eui.eu
cadmus.eui.eu

Estimation methods comparison of SVAR model with the mixture of two normal distributions - Monte Carlo analysis

Katarzyna Maciejowska
EUI

June 6, 2010

Abstract

This paper addresses the issue of obtaining maximum likelihood estimates of parameters for structural VAR models with a mixture of distributions. Hence the problem does not have a closed form solution, numerical optimization procedures need to be used. A Monte Carlo experiment is design to compare the performance of four maximization algorithms and two estimation strategies. It is shown that the EM algorithm outperforms the general maximization algorithms such as BFGS, NEWTON and BHHH. Moreover simplification of the probelm introduced in the two steps quasi ML method does not worsen small sample properties of the estimators and therefore may be recommended in the empirical analysis.

JEL classification: C32, C46
Keywords: Structural vetcor autoregression , Error correction models, Mixed normal, Monte Carlo

1 Introduction

Structural vector autoregressive (SVAR) models are widely used in applied macroeconomics. They allow for the estimation of structural shocks and impulse responses from empirical data and therefore, can be used to evaluate economic theory. However, this class of models requires additional information about the theoretical setup or the data in order to identify the structural parameters. A standard approach to obtain identifiability is to impose parameter constraints that can be justified by the economic theory. Unfortunately, there is no agreement on which of the identification schemes should be used and imposing just-identifying restrictions makes it impossible to empirically evaluate some of the underlying economic assumptions. The above critique raises the question of whether there is a property of the data instead of the economic theory that can be used to identify SVAR parameters. Rigobon (2003) shows that if there is a shift in the variance of the structural shocks it can provide enough information to identify the SVAR model. Lanne and Lütkepohl (2008) generalizes this approach and develops a test for the presence of a variance shift and for the stability of the correlation structure. This paper follows the specification of Lanne and Lütkepohl (2005), which assumes nonnormality of structural shocks rather then a discrete change in the variance. The residuals are allowed to be distributed according to the mixture of two normal distributions and it is demonstrated how this property can be used to identify the parameters.

Scientific literature provides many works that discuss the issue of mixture models. Mixture models can be found both in economics and in other disciplines such as biology, medicine, engineering and marketing, among others. They were first used by biometrician Karl Pearson (1894), who analyzed a population of crabs and proved the existence of two subspecies in the examined sample. In the 1960s economists tried to use the ML approach to estimate the model parameters (Day (1969)). However, it was the EM algorithm described by Dempster, Laird, and Rubin (1977) that significantly simplified the estimation procedure and therefore helped to popularize the mixture models.

The mixture models are also special cases of Markov switching (MS) models. A Markov process simplifies to a mixture distribution if diagonal elements of its transition matrix sum to one. Markov switching models are very flexible and can account for both nonliearities in the mean and heteroscedasticity. They are extensively used in econometrics (Kim and Nelson (1999), Sims and Zha (2006), Smith, Naik, and Tsai (2006)), especially in business cycle analysis (Hamilton (1989), Goodwin (1993), Diebold and Rudebusch (1996), Kim and Nelson (1998)). They were popularized by the seminal paper Hamilton (1989), which discusses the estimation issues for univariate processes. The approach was extended to a multivariate case by Krolzig (1997).

An open question that still needs to be examined are small sample properties of mixture model estimators. This issue is of special interest when mixture models are applied in macroeconomic analysis because they are associated with a usage of relatively short time series. Therefore, the main scope of the paper is to evaluate the performance of different estimation methods and maximization algorithms in the context of SVAR models with mixtures of normal distributions, as proposed by Lanne and Lütkepohl (2005), and discuss the difficulties associated with the estimation process. Since the mixture models are special cases of MS models, we believe that our research also contributes to the discussion on estimation issues of MS models, especially in the context of structural analysis.

The paper is structured as follows. In Section 2, SVAR model with a mixture of two normal distributions is introduced and the identification issues are discussed. Estimation methods and optimization algorithms are considered in Section 3. In Section 4, a Monte Carlo experiment is described and results for different estimation methods and optimization algorithms are presented. Finally, conclusions are provided in Section 5.

2 SVAR models with a mixture of normal distributions

2.1 Model description

The literature discusses different types of SVAR models: A-model, B-model and AB-model (see Lütkepohl (2005)). The classification depends on the relationships the model attempt to describe, i.e., whether we are interested in the relations between the observable variables or responses to unobservable impulses. In this paper we will focus on the B-model that describes the direct, instantaneous effect of the structural shocks on the endogenous variables. In the B-model it is assumed that the forecast error ε is a linear function of the structural shock, u. The model can be written in the following way

$$y_t = A_0 + \sum_{i=1}^{p} A_i y_{t-i} + \varepsilon_t \tag{1}$$

where $\varepsilon_t = Bu_t$ and the variance-covariance matrices of structural and forecast errors are $\Sigma_u = I_k$ and $\Sigma_\varepsilon = BB'$, respectively.

In the setup, y_t is a $k \times 1$ vector of endogenous variables, ε_t is a $k \times 1$ vector of forecast errors and u_t is a $k \times 1$ vector of structural shocks with an identity covariance matrix $\Sigma_u = I_k$. A_0 is a $k \times 1$ vector of constants and $A_i, i = 1, ..., p$ are $k \times k$ matrices of the autoregressive parameters. B is a $k \times k$ nonsingular matrix that describes the transition mechanisms of the structural shocks u_t.

The structural VAR model has $k + p \cdot k^2 + k^2$ unknown parameters. The reduced form of the model (1) allows for estimation of only $k + p \cdot k^2 + k\left(k+1\right)/2$ parameters. In order to identify all structural parameters, an additional $k\left(k-1\right)/2$ linearly independent restrictions need to be imposed.

Lanne and Lütkepohl (2005) proposes solving the identification problem by making an assumption on the distribution of shocks. It is assumed that the structural shocks vector, u_t, has a mixed normal distribution. It means that

$$u_t \sim \begin{cases} N(0,I_k) & \text{with probability } \gamma \\ N(0,\Psi) & \text{with probability } 1-\gamma \end{cases}$$

where the variance-covariance matrix Ψ is diagonal. Under this specification, the unconditional variance of the structural shock is $\Sigma_u = \gamma I_k + (1-\gamma)\Psi$. The matrix Σ_u is no longer identity matrix but it is still diagonal. The diagonality of the matrix Σ_u ensures that the structural shocks are uncorrelated. Lanne and Lütkepohl (2005) proves that if all diagonal elements of the matrix Ψ are distinct then the structural parameters of the model are identifiable. The issue of identifiability will be discussed in more detail in Section 2.3.

2.2 Density function of forecast errors

In order to analyze the properties of the model we need to derive the density function for the forecast errors. Since the errors, ε_t, are a linear combination of the structural shocks, u_t, then they also have a mixed normal distribution

$$\varepsilon_t \sim \begin{cases} N(0,BB') & \text{with probability } \gamma \\ N(0,B\Psi B') & \text{with probability } 1-\gamma \end{cases}$$

Therefore, the density function $f(\varepsilon_t; B, \Psi, \gamma)$ is given by

$$\begin{aligned} f(\varepsilon_t; B, \Psi, \gamma) \quad = \quad & \gamma (2\pi)^{-k/2} \det (BB')^{-1/2} \exp\left(-\frac{1}{2}\varepsilon_t' (BB')^{-1} \varepsilon_t\right) \\ & + (1-\gamma)(2\pi)^{-k/2} \det (B\Psi B')^{-1/2} \exp\left(-\frac{1}{2}\varepsilon_t' (B\Psi B')^{-1} \varepsilon_t\right) \end{aligned} \tag{2}$$

The function is a sum of two components

$$f(\varepsilon_t; B, \Psi, \gamma) = \gamma f_1(\varepsilon_t; B) + (1-\gamma) f_2(\varepsilon_t; B, \Psi) \tag{3}$$

where

$$f_1(\varepsilon_t; B) = (2\pi)^{-k/2} \det (BB')^{-1/2} \exp\left(-\frac{1}{2}\varepsilon_t' (BB')^{-1} \varepsilon_t\right)$$

and

$$f_2(\varepsilon_t; B, \Psi) = (2\pi)^{-k/2} \det (B\Psi B')^{-1/2} \exp\left(-\frac{1}{2}\varepsilon_t' (B\Psi B')^{-1} \varepsilon_t\right)$$

Under the assumption of no time correlation of errors, the joint density can be written as follows

$$f(\varepsilon; B, \Psi, \gamma) = \prod_{t=1}^{T} f(\varepsilon_t; B, \Psi, \gamma)$$

with $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T\}$.

In further sections, for notational simplicity, $f(\varepsilon_t; \theta, \gamma)$ is used instead of $f(\varepsilon_t; B, \Psi, \gamma)$, where $\theta' = \{vec(B)' : diag(\Psi)'\}$.

2.3 Identification

There is a theoretical question whether it is possible to uniquely identify the parameters of SVAR models with the mixture of two normal distributions. In the literature there are papers that address the issue of parameters identification in different kinds of models. Following Rothenberg (1971), we can distinguish between locally and globally identifiable structures. Let us denote $f(\varepsilon; \delta)$ as a density function of a random variable ε for parameters $\delta \in \Delta$.

Definition 1 *A parameter point $\delta \in \Delta$ is said to be globally identifiable if there is no other $\tilde{\delta} \in \Delta$ such that $f\left(\varepsilon; \tilde{\delta}\right) = f(\varepsilon; \delta)$ for all ε.*

Definition 2 *A parameter point $\delta \in \Delta$ is said to be locally identifiable if there exists an open neighborhood of δ containing no other $\tilde{\delta}$ such that $f\left(\varepsilon; \tilde{\delta}\right) = f(\varepsilon; \delta)$ for all ε.*

In the case of standard mixture models, it is straightforward to see that they are not globally identifiable. One can always change the order of the mixture components without changing the overall distribution. This problem is known as the "label switching". In the simple mixture model, in which the density function is described by

$$f(\varepsilon; \theta, \gamma) = \sum_{i=1}^{n} \gamma_i f_i(\varepsilon; \theta_i)$$

where $\theta = \{\theta_1, ..., \theta_n\}$ is a set of mixture components parameters and $\gamma = \{\gamma_1, ..., \gamma_n\}$ is a set of mixing proportions, such that for all $i \in \{1, .., n\}$ $\gamma_i > 0$ and $\sum_{i=1}^{n} \gamma_i = 1$, "label switching" means that for any permutation of indices $k_1, ..., k_n$

$$f\left(\varepsilon; \tilde{\theta}, \tilde{\gamma}\right) = \sum_{i=1}^{n} \gamma_{k_i} f_{k_i}(\varepsilon; \theta_{k_i}) = \sum_{i=1}^{n} \gamma_i f_i(\varepsilon; \theta_i) = f(\varepsilon; \theta, \gamma)$$

where $\tilde{\theta} = \{\theta_{k_i}, ..., \theta_{k_n}\}$ and $\tilde{\gamma} = \left\{\gamma_{k_1}, ..., \gamma_{k_n}\right\}$.

In the SVAR model with the mixture of two normal distributions, the error term ε_t follows (3). It means that the mixture components are defined by different parameter vectors. Thus, components cannot be simply flipped around by changing their order. However, for any B, Ψ and γ, there exist $\tilde{B} = B\Psi^{0.5}$, $\tilde{\Psi} = \Psi^{-1}$ and $\tilde{\gamma} = 1 - \gamma$ such that for all $\varepsilon \in R$ there is $f\left(\varepsilon; \tilde{B}, \tilde{\Psi}, \tilde{\gamma}\right) = f(\varepsilon; B, \Psi, \gamma)$. The proof can be found in Appendix 1.6.1.

An additional problem that arises from the specification of SVAR models is the identifiability of the matrices B and Ψ. It can be shown that one can change the order of columns of B and corresponding diagonal elements of Ψ without influencing the values of the likelihood function. Moreover, the columns of B can be multiplied by -1 and it will not affect the values of the density function.

There are no doubts that the parameters of the SVAR models with the mixture of two normal distributions are not globally identifiable. It was shown, however, by Lanne and Lütkepohl (2005) that under some mild conditions they may be locally identifiable. The necessary and sufficient condition for the local identification is that the diagonal elements of the matrix Ψ are all mutually different.

3 Estimation methods

The problem of estimating parameters of mixture models has been a subject of a large body of literature. Redner and Walker (1984) and McLachlan and Peel (2000) provide a survey of both theoretical and empirical publications discussing the properties and applications of different types of estimators. Recently, due to the increase of computational efficiency, most of the research concentrates on the application of the maximum-likelihood method. As the functional form of the residual distribution in the mixture models is usually treated as known, ML seems to be a plausible approach.

In the presented work, two estimation methods will be used. First, the standard maximum likelihood estimation will be described. Second, a two steps quasi ML estimation, which allows for the estimation of the autoregressive and mixture parameters separately, will be presented. Finally, the properties of the ML estimators will be discussed.

3.1 Maximum Likelihood and two steps quasi Maximum Likelihood estimators

The maximum likelihood estimation method depends on the assumed functional form of the joint error distribution. In the SVAR model with the mixture of two normal densities, the p.d.f. of the forecast errors, ε_t, for a given period t is given by (2). Therefore, the value of the log-likelihood function $L(\theta, \gamma|\varepsilon_t)$ for the t-th error, ε_t, is

$$\begin{aligned} L(\theta,\gamma|\varepsilon_t) &= \ln\left(f\left(\varepsilon_t;\theta,\gamma\right)\right) \\ &= -\frac{k}{2}\ln(2\pi) \\ &\quad + \ln\left(\begin{array}{c} \gamma \det\left(BB'\right)^{-1/2}\exp\left(-\frac{1}{2}\varepsilon_t'\left(BB'\right)^{-1}\varepsilon_t\right) + \\ (1-\gamma)\det\left(B\Psi B'\right)^{-1/2}\exp\left(-\frac{1}{2}\varepsilon_t'\left(B\Psi B'\right)^{-1}\varepsilon_t\right) \end{array}\right) \end{aligned}$$

A constant term $-\frac{k}{2}\log(2\pi)$ will be omitted in further analysis. The joint log-likelihood is

$$\begin{aligned} L(\theta,\gamma|\varepsilon) &= \sum_{t=1}^{T} L(\theta,\gamma|\varepsilon_t) \\ &= \ln\left(f\left(\varepsilon_t;\theta,\gamma\right)\right) \end{aligned}$$

The maximization problem

$$\max_{\theta\in\Omega,\gamma\in(0,1)} L(\theta,\gamma|\varepsilon) = \max_{\theta\in\Omega,\gamma\in(0,1)} \sum_{t=1}^{T} \ln\left(f\left(\varepsilon_t;\theta,\gamma\right)\right)$$

where θ is a vector of parameters defined as before and

$$\Omega = \{\theta : \det(B) \neq 0, diag(\Psi) > 0\}$$

is a set of all possible parameter vectors, does not have a closed form solution and therefore iterative optimization procedures have to be used.

3.1.1 One step Maximum Likelihood

In this method one searches for the maximum of the log-likelihood function over both the autoregressive and mixture parameters. We can rewrite the model with the lag polynomial

$$A(L)y_t - A_0 = \varepsilon_t$$

where $A(L) = I_k - \sum_{i=1}^{p} A_i L^i$ and L is a lag operator, such that $L^i y_t = y_{t-i}$. A_0 is a $k \times 1$ vector of constants. Then the estimators $\hat{A}_{0,}\hat{A}_1, ..., \hat{A}_p, \hat{B}, \hat{\Psi}, \hat{\gamma}$ are chosen to maximize

$$L(\theta,\gamma,A|y) = \sum_{t=p}^{T} \ln f(A(L)y_t - A_0;\theta,\gamma)$$

where $A = (A_0, A_1, ..., A_p)$, $y = (y_{1,}y_2, ..., y_T)$ and $f(.;\theta,\gamma)$ is defined in (2).

3.1.2 Two steps quasi Maximum Likelihood

In this method the estimation procedure consists of two steps. Firstly, the autoregressive parameters are estimated with the LS or quasi ML method. Then the estimates of the residuals are computed according to the formula

$$\hat{e}_t = y_t - \left(\hat{A}_0 + \sum_{i=1}^{p} \hat{A}_i y_{t-i}\right)$$

Finally, the mixture of two normal distributions is fitted to the estimated residuals $\hat{e}_t$ with the ML method. Then parameters $\hat{B}, \hat{\Psi}, \hat{\gamma}$ are chosen to maximize

$$L(\theta,\gamma|\hat{e}) = \sum_{t=p}^{T} \ln f(\hat{e}_t;\theta,\gamma)$$

where $\hat{e} = (\hat{e}_p, \hat{e}_{p+1}, ..., \hat{e}_T)$ and $f(.)$ is defined as in (2).

This is a quasi ML method because it is conditional on the estimates of the estimates of the autoregressive parameters, which in principle differs form the true ones. Thus,

$$L(\theta, \gamma|\hat{e}) \neq L(\theta, \gamma|\varepsilon)$$

Fortunately, the autoregressive parameters can be consistently estimated with the LS or quasi ML method and therefore, the estimates of the mixture parameters $\hat{B}$, $\hat{\Psi}$ and $\hat{\gamma}$ converge to the true ones. This estimation method is however less efficient than the full Maximum Likelihood approach.

3.2 Numerical maximization algorithms

As mentioned before, the ML problem does not have a closed form solution. Therefore, numerical maximization algorithms need to be used to obtain the ML estimates of the parameters. There exist general iterative procedures, such as Newton's methods, two steps quasi Newton's methods and conjugate gradient methods, which can be used in this context. There are, however, other methods that are more specific and thus more suitable for the mixture distributions models. One of them is the EM algorithm. It was formalized by Dempster, Laird, and Rubin (1977) and designed for estimation problems with incomplete data. McLachlan and Krishnan (1997) provides a broad review of the literature dedicated to its theoretical and empirical properties.

3.2.1 EM algorithm

The estimation of the SVAR models with the mixture of distributions can be analyzed from the perspective of the incomplete data problem. Let us assume that the data generating process of the shocks ε_t is

$$\varepsilon_t \sim \begin{cases} N(0{,}BB') & \text{if } \mathrm{Z}_t = 1 \\ N(0{,}B\Psi B') & \text{if } \mathrm{Z}_t = 0 \end{cases}$$

where Z_t is an indicator variable. Then the density function of ε_t conditional on Z_t could be rewritten as follows

$$f(\varepsilon_t|Z_t;\theta) = f_1(\varepsilon_t;\theta)^{Z_t} f_2(\varepsilon_t;\theta)^{1-Z_t}$$

where $f_1(\varepsilon_t;\theta)$ and $f_2(\varepsilon_t;\theta)$ are defined in Section 2.2.

In the mixture model the mixing probabilities are assumed to be constant over time. It corresponds to the assumption

$$\begin{aligned} prob(Z_t = 1) &= \gamma \\ prob(Z_t = 0) &= 1 - \gamma \end{aligned}$$

Therefore, Z_t needs to have a Bernoulli distribution

$$g\left(Z_t;\gamma\right)=\gamma^{Z_t}\left(1-\gamma\right)^{1-Z_t}$$

The joint density function of ε_t and Z_t is given by

$$f_c\left(\varepsilon_t,Z_t;\theta,\gamma\right)=f_1\left(\varepsilon_t;\theta\right)^{Z_t}f_2\left(\varepsilon_t;\theta\right)^{1-Z_t}\gamma^{Z_t}\left(1-\gamma\right)^{1-Z_t}$$

and

$$\begin{aligned}\ln\left(f_c\left(\varepsilon_t,Z_t;\theta,\gamma\right)\right) &= Z_t\left\{\ln\left(\gamma\right)+\ln\left(f_1\left(\varepsilon_t;\theta\right)\right)\right\}\\ &\quad+\left(1-Z_t\right)\left\{\ln\left(1-\gamma\right)+\ln\left(f_2\left(\varepsilon_t;\theta\right)\right)\right\}\end{aligned}$$

The complete-data log likelihood $L_c\left(\theta,\gamma|\varepsilon\right)$ (meaning that both ε_t and Z_t are assumed to be observable) can be written as follows

$$\begin{aligned}L_c\left(\theta,\gamma|\varepsilon\right) &= \sum_{t=1}^{T}L_c\left(\theta,\gamma|\varepsilon_t\right)\\ &= \sum_{t=1}^{T}\ln\left(f_c\left(\varepsilon_t,Z_t;\theta,\gamma\right)\right)\end{aligned}$$

Therefore,

$$\begin{aligned}L_c\left(\theta,\gamma|\varepsilon\right) &= \sum_{t=1}^{T}Z_t\left\{\ln\left(\gamma\right)+\ln\left(f_1\left(\varepsilon_t;\theta\right)\right)\right\}\\ &\quad+\sum_{t=1}^{T}\left(1-Z_t\right)\left\{\ln\left(1-\gamma\right)+\ln\left(f_2\left(\varepsilon_t;\theta\right)\right)\right\}\end{aligned}$$

The EM algorithm consists of two steps: E (computing the expectation of $L_c\left(\theta,\gamma|\varepsilon\right)$ conditional on the the observable data ε_t) and M (maximizing the expected $L_c\left(\theta,\gamma|\varepsilon\right)$ over the parameter space $\Omega\cup(0,1)$.

E - Step In this step the expected value of the complete-data log likelihood is computed. The expected value of the $L_c\left(\theta,\gamma|\varepsilon\right)$ conditional on the the observable data ε for an initial parameters vector θ_0 and γ_0 is given by $Q\left(\theta,\gamma;\theta_0,\gamma_0\right)$

$$
\begin{aligned}
Q\left(\theta,\gamma;\theta_0,\gamma_0\right) &= E\left(\sum_{t=1}^{T} Z_t\left\{\log\left(\gamma\right)+\log\left(f_1\left(\varepsilon_t;\theta\right)\right)\right\}|\varepsilon;\theta_0,\gamma_0\right)\\
&\quad +E\left(\sum_{t=1}^{T}\left(1-Z_t\right)\left\{\log\left(1-\gamma\right)+\log\left(f_2\left(\varepsilon_t;\theta\right)\right)\right\}|\varepsilon;\theta_0,\gamma_0\right)\\
&= \sum_{t=1}^{T} E\left(Z_t|\varepsilon;\theta_0,\gamma_0\right)\left\{\log\left(\gamma\right)+\log\left(f_1\left(\varepsilon_t;\theta\right)\right)\right\}\\
&\quad +\sum_{t=1}^{T} E\left(1-Z_t|\varepsilon;\theta_0,\gamma_0\right)\left\{\log\left(1-\gamma\right)+\log\left(f_2\left(\varepsilon_t;\theta\right)\right)\right\}
\end{aligned}
$$

Let us denote by $\tau_t\left(\theta_0,\gamma_0\right)$ an expected value of the indicator variable Z_t for the initial parameters values θ_0 and γ_0

$$
\begin{aligned}
\tau_t\left(\theta_0,\gamma_0\right) &= E\left(Z_t|\varepsilon;\theta_0,\gamma_0\right)\\
&= 0\cdot f\left(Z_t=0|\varepsilon;\theta_0,\gamma_0\right)+1\cdot f\left(Z_t=1|\varepsilon;\theta_0,\gamma_0\right)\\
&= f_c\left(\varepsilon_t,Z_t=1;\theta_0,\gamma_0\right)/f\left(\varepsilon_t;\theta_0\right)\\
&= \gamma_0 f_1\left(\varepsilon_t;\theta_0\right)/f\left(\varepsilon_t;\theta_0\right)
\end{aligned}
$$

Then

$$
\begin{aligned}
E\left(1-Z_t|\varepsilon;\theta_0,\gamma_0\right) &= 1-\gamma_0 f_1\left(\varepsilon_t;B_0\right)/f\left(\varepsilon_t;\theta_0\right)\\
&= 1-\tau_t\left(\theta_0,\gamma_0\right)
\end{aligned}
$$

Thus, $Q\left(\theta,\gamma;\theta_0,\gamma_0\right)$ takes the form

$$
\begin{aligned}
Q\left(\theta,\gamma;\theta_0,\gamma_0\right) &= \sum_{t=1}^{T}\tau_t\left(\theta_0,\gamma_0\right)\left\{\log\left(\gamma\right)+\log\left(f_1\left(\varepsilon_t;\theta\right)\right)\right\}\\
&\quad +\sum_{t=1}^{T}\left(1-\tau_t\left(\theta_0,\gamma_0\right)\right)\left\{\log\left(1-\gamma\right)+\log\left(f_2\left(\varepsilon_t;\theta\right)\right)\right\}
\end{aligned}
$$

M - Step In this step the new estimates of θ and γ are chosen to maximize $Q\left(\theta,\gamma;\theta_0,\gamma_0\right)$.

$$
\left(\hat{\theta},\hat{\gamma}\right)=\arg\max_{\theta\in\Omega,\gamma\in(0,1)} Q\left(\theta,\gamma;\theta_0,\gamma_0\right)
$$

The $Q\left(\theta,\gamma;\theta_0,\gamma_0\right)$ function can be decomposed into two parts

$$
Q\left(\theta,\gamma;\theta_0,\gamma_0\right)=Q_1\left(\gamma;\theta_0,\gamma_0\right)+Q_2\left(\theta;\theta_0,\gamma_0\right)
$$

such that

$$\begin{aligned} Q_1(\gamma;\theta_0,\gamma_0) &= \log(\gamma)\sum_{t=1}^{T}\tau_t(\theta_0,\gamma_0)+\log(1-\gamma)\sum_{t=1}^{T}\{1-\tau_t(\theta_0,\gamma_0)\} \\ &= \log(\gamma)\sum_{t=1}^{T}\tau_t(\theta_0,\gamma_0)+\log(1-\gamma)\left\{T-\sum_{t=1}^{T}\tau_t(\theta_0,\gamma_0)\right\} \end{aligned}$$

$$Q_2(\theta;\theta_0,\gamma_0)=\sum_{t=1}^{T}\tau_t(\theta_0,\gamma_0)\log(f_1(\varepsilon_t;\theta))+(1-\tau_t(\theta_0,\gamma_0))\log(f_2(\varepsilon_t;\theta))$$

The first component depends only on the mixing proportions γ whereas the second one depends on θ. Consequently, the maximization problem can be solved by separately estimating the proportion parameter γ and the rest of the parameters θ. It can be easily shown that the $Q_1(\gamma;\theta_0,\gamma_0)$ is maximized by

$$\hat{\gamma}=\sum_{t=1}^{T}\tau_t(\theta_0,\gamma_0)/T$$

Finally,

$$\hat{\theta}=\arg\max_{\theta\in\Omega}Q_2(\theta;\theta_0,\gamma_0)$$

Iterations of the algorithm Once the new estimates of the parameters $\hat{\theta}$ and $\hat{\gamma}$ are obtained, the two steps E and M are repeated for $\theta_0=\hat{\theta}$ and $\gamma_0=\hat{\gamma}$. The algorithm is terminated when a stopping condition is fulfilled. There are two popular stopping rules

1. The algorithm is stopped when the value of the log-likelihood function does not change by more then δ

$$\left|\log L\left(\hat{\theta},\hat{\gamma}|\varepsilon\right)-\log L(\theta_0,\gamma_0|\varepsilon)\right|\leq\delta$$

2. The algorithm is stopped when the parameters do not change much. It means that for some chosen δ

$$\left\|\bar{\theta}-\bar{\theta}_0\right\|\leq\delta$$

where $\|.\|$ denotes some norm and $\bar{\theta}=\left(\hat{\theta}',\hat{\gamma}\right)'$, $\bar{\theta}_0=\left(\theta_0',\gamma_0\right)'$.

3.3 Problems with Maximum Likelihood estimation

The maximum likelihood estimators suffer from two problems: the likelihood function is unbounded and the parameters are not globally identified. The second issue was discussed before and due to local identifiability does not threaten the estimation process but influences the interpretation of the estimated parameters. The first one is much more serious and some modification of the estimation procedures need to be considered.

3.3.1 Unbounded Likelihood function

An example of an unbounded likelihood function for a mixture model was given by Kiefer and Wolfowitz (1956). Let us consider an univariate, mixture model with a shift in a variance

$$x_t \sim \begin{cases} N(\mu{,}1) & \text{with probability } 0.5 \\ N(\mu,\sigma^2) & \text{with probability } 0.5 \end{cases}$$

Then the density function for x_t is given by

$$\begin{aligned} f\left(x_t;\mu,\sigma\right) &= 0.5\frac{1}{\left(2\pi\right)^{0.5}}\exp\left(-0.5\left(x_t-\mu\right)^2\right) \\ &\quad +0.5\frac{1}{\left(2\pi\right)^{0.5}}\frac{1}{\sigma}\exp\left(-0.5\frac{\left(x_t-\mu\right)^2}{\sigma^2}\right) \end{aligned}$$

Let us assume that there is a finite number of observations $\{x_t\}$ and $\max_t|x_t-\mu| = m < \infty$. Suppose we choose $\mu = x_1$ and a sequence of standard deviations $\sigma_n \to 0$. Then, for all $x_t = \mu$, the density function diverges to infinity.

$$\begin{aligned} f\left(x_t;\mu,\sigma_n\right) &= 0.5\frac{1}{\left(2\pi\right)^{0.5}}\exp\left(-0.5\left(x_t-\mu\right)^2\right) \\ &\quad +0.5\frac{1}{\left(2\pi\right)^{0.5}}\frac{1}{\sigma_n}\exp\left(-0.5\frac{\left(x_t-\mu\right)^2}{\sigma_n^2}\right) \\ &= 0.5\frac{1}{\left(2\pi\right)^{0.5}}+0.5\frac{1}{\left(2\pi\right)^{0.5}}\frac{1}{\sigma_n}\to\infty \end{aligned}$$

The density for $x_t \neq \mu$ is bounded away from zero

$$
\begin{aligned}
f\left(x_t;\mu,\sigma_n\right) &= 0.5\frac{1}{(2\pi)^{0.5}}\exp\left(-0.5\left(x_t-\mu\right)^2\right) \\
&\quad +0.5\frac{1}{(2\pi)^{0.5}}\frac{1}{\sigma_n}\exp\left(-0.5\frac{\left(x_t-\mu\right)^2}{\sigma_n^2}\right) \\
&\to 0.5\frac{1}{(2\pi)^{0.5}}\exp\left(-0.5\left(x_t-x_1\right)^2\right) \\
&\geq 0.5\frac{1}{(2\pi)^{0.5}}\exp\left(-0.5m^2\right)>0
\end{aligned}
$$

Thus, $L\left(\mu,\sigma_n|x\right)=\prod_{t=1}^{T}f\left(x_t;\mu,\sigma_n\right)\to\infty$

The problem seems to be equally severe for the SVAR models with a mixture of two normal distributions. The density function for an error, ε_t, is given by the following formula

$$
\begin{aligned}
f\left(\varepsilon_t\right) &= \gamma\left(2\pi\right)^{-k/2}\det\left(B\right)^{-1}\exp\left(-\frac{1}{2}\left(B^{-1}\varepsilon_t\right)'B^{-1}\varepsilon_t\right)+ \\
&\left(1-\gamma\right)\left(2\pi\right)^{-k/2}\det\left(\Psi\right)^{-1/2}\det\left(B\right)^{-1}\exp\left(-\frac{1}{2}\left(B^{-1}\varepsilon_t\right)'\Psi^{-1}B^{-1}\varepsilon_t\right)
\end{aligned}
$$

We can always find a matrix B such that $\det\left(B\right)<M_1<\infty$ and there exists a time index $s\in\{1,...,T\}$ such that the ith element of $b_s=B^{-1}\varepsilon_s$ is equal to zero, $b_{is}=\left[B^{-1}\varepsilon_s\right]_i=0$, for some $i\in\{1,...,k\}$. We can choose the sequence Ψ_n of diagonal, positive definite matrices that satisfies $\Psi_{ii}^n\to 0$ and $\Psi_{jj}^n>M_2>0$ for $j\neq i$. We know that

$$
-\left(B^{-1}\varepsilon_t\right)'\Psi^{-1}B^{-1}\varepsilon_t=-\sum_{j\neq i}\frac{1}{\Psi_{jj}}b_{jt}^2-\frac{1}{\Psi_{ii}}b_{it}^2
$$

For $t=s$

$$
\frac{1}{\Psi_{ii}}b_{it}^2=0
$$

Therefore,

$$
-\left(B^{-1}\varepsilon_t\right)'\Psi^{-1}B^{-1}\varepsilon_t=-\sum_{j\neq i}\frac{1}{\Psi_{jj}^n}b_{jt}^2>-\frac{1}{M_2}\sum_{j\neq i}b_{jt}^2>-\infty
$$

and

$$
\exp\left(-\frac{1}{2}\left(B^{-1}\varepsilon_t\right)'\Psi_n^{-1}B^{-1}\varepsilon_t\right)\gg 0
$$

Since

$$\det(\Psi_n) \to 0$$

then

$$\det(\Psi_n)^{-1/2} \det(B)^{-1} \exp\left(-\frac{1}{2}\left(B^{-1}\varepsilon_t\right)' \Psi^{-1} B^{-1}\varepsilon_t\right) \to \infty.$$

Thus, $f(\varepsilon_t) \to \infty$.

For $t \neq s$ the value of density function $f(\varepsilon_t)$ is bounded away from zero

$$f(\varepsilon_t) > \gamma (2\pi)^{-k/2} \det(B)^{-1} \exp\left(-\frac{1}{2}\left(B^{-1}\varepsilon_t\right)' B^{-1}\varepsilon_t\right) > 0$$

So $L(\theta, \gamma|\varepsilon) = \prod_{t=1}^{T} f(\varepsilon_t) \to \infty$. Therefore the likelihood function is unbounded.

The problem of an unbounded likelihood function rises some questions about the ML estimators.

What is the ML estimator for the unbounded likelihood function? When the likelihood function is unbounded then the global maximizer of the likelihood function does not exist. Therefore one can not talk about the ML estimator in the traditional sense (see McLachlan and Peel (2000) for some discussion). It does not mean, however, that there is no sequence of local maximizers with properties of consistency, efficiency and asymptotic normality. Redner and Walker (1984) provides the regularity conditions under which, for the class of locally identifiable mixtures, such a sequence exists. Moreover, when the parameter space is compact and contains the true parameters in its interior, the MLE is a point at which the likelihood obtains its largest local maximum.

How can the ML estimation procedure be improved? Hathaway (1985) proposes imposing a set of constraints (ensuring that the parameter space is compact and does not include singularity points) that allows for the consistent estimation of the parameters. In the case of univariate time series, the constraint is $\min_{i,j}(\sigma_i/\sigma_j) \geq c$ for some constant $c > 0$. In the multivariate case, Hathaway (1985) proposes to constrain all of the characteristic roots of $\Sigma_i \Sigma_j^{-1}$ (for any $1 \leq i \neq j \leq k$) to be greater or equal to some minimum value $c > 0$. These kind of restrictions will lead to constrained (global) maximum-likelihood formulations which are strongly consistent (if they are satisfied by the true parameters). The main disadvantage of the approach is the arbitrary choice of the value of $c > 0$. It is particularly difficult, when there is no initial intuition about the data generating process and no information to base the guess on.

Some other forms of the constraints are discussed in the literature. For example, McLachlan and Peel (2000) proposes to limit the distance between the component generalized variances by restricting the ratio $|\Sigma_i| / |\Sigma_j|$ to be greater or equal to $c > 0$.

What can we do in the case of the SVAR models with mixture of two normal densities? One may want to impose similar constraints on the parameters in the case of the SVAR model with the mixture of two normal densities. There are, however, differences between the setup presented in this paper and one discussed typically in the literature, they are associated with the components variances. In the SVAR models, the variances are composed of two matrices: B and Ψ: $\Sigma_1 = BB'$ and $\Sigma_2 = B\Psi B'$. Thus

$$\Sigma_2\Sigma_1^{-1} = B\Psi B' \cdot B'^{-1}B^{-1} = B\Psi B^{-1}$$

Let us denote by $\lambda\left(A\right)$ a set of all eigenvalues of the square matrix A. Then

$$\lambda\left(\Sigma_2\Sigma_1^{-1}\right) = \lambda\left(B\Psi B^{-1}\right) = \lambda\left(\Psi\right) = diag\left(\Psi\right)$$

So the Hathaway constraints for the two components case are equivalent to the following

$$\begin{array}{l} 0 < c \leq \min_{i\in\{1,..,K\}} \Psi_{i,i} \\ \max_{i\in\{1,..,K\}} \Psi_{i,i} \leq 1/c < \infty \end{array} \tag{4}$$

How to treat the obtained results? How can we evaluate the local maximum we find? The mixture models suffer not only the problem of unbounded likelihood function but also the problem of spurious maximizers. Spurious maximizers are typically generated by a small group of observations, which are located close together (Day (1969)). They are characterized by a big relative difference between components variances. Thus imposing restrictions on the parameters may reduce the number of spurious maximizers. The minimum eigenvalue of the $\Sigma_i\Sigma_j^{-1}$ can also be used to evaluate the local maximizers of the unconstrained likelihood and to choose the most interesting one.

4 Monte Carlo Experiment

The purpose of the Monte Carlo experiment is to investigate how a choice of an estimation method and maximization algorithm influences estimates of the parameters. The exercise helps to answer the question what is the cost of using the two steps quasi ML instead of ML method. If there are no significant differences, then the two steps quasi ML approach will be a very attractive from the practical point of view as it allows to reduce significantly the complexity of the problem. Other interesting issues are the ability of different maximization algorithms to find the true, rather than spurious, local maximizers and the robustness to the guesses of the initial parameter values.

4.1 Experimental design

In the experiment, two data generating processes are considered: VAR in levels and VECM, both with the mixture of two normal distributions. The VECM process

$$\Delta y_t = A_0 + \alpha\beta' y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + Bu_t$$

can be represented as a VAR process

$$y_t = A_0 + \sum_{j=1}^{p} A_j y_{t-j} + Bu_t$$

where the relationship between the VECM and VAR parameters is described as follow:

$$\begin{array}{rcl} A_1 & = & \alpha\beta' + \Gamma_1 + I_k \\ A_2 & = & \Gamma_2 - \Gamma_1 \\ & & \vdots \\ A_{p-1} & = & \Gamma_{p-1} - \Gamma_{p-2} \\ A_p & = & -\Gamma_{p-1} \end{array}$$

Therefore, in both cases the data sets used in the research can be generated according to the VAR specification. It is assumed that u_t follows a mixture of two normal distributions $N(0, I)$ and $N(0, \Psi)$ with mixing proportions γ and $1 - \gamma$ ($\gamma \in (0, 1)$), respectively

For each type of data generating process, the Monte Carlo experiment consists of 1000 replications. In each replication ($i = 1, \dots 1000$), a time series is generated according to the following algorithm :

1. For each replication i and time period t a variable Z_{it} is generated from the binomial distribution with $prob(Z_{it} = 1) = \gamma$ and $prob(Z_{it} = 0) = 1 - \gamma$. Firstly, we draw randomly v_{it} form the uniform distribution on the interval $[0, 1]$. Then the value of Z_{it} is assigned $Z_{it} = 1 \Leftrightarrow v_{it} \leq \gamma$, $Z_{it} = 0 \Leftrightarrow v_{it} > \gamma$.

2. Structural shocks u_{it} are generated according to the distribution $N(0, I)$ if $Z_{it} = 1$ and $N(0, \Psi)$ if $Z_{it} = 0$ for each time period t (or alternatively $u_{it} \sim N(0, I) \Leftrightarrow v_{it} \leq \gamma$, $u_{it} \sim N(0, \Psi) \Leftrightarrow v_{it} > \gamma$).

3. Time series $\{y_{it}\}$ are generated from the formula

$$y_{it} = A_0 + \sum_{j=1}^{p} A_p y_{i,t-j} + Bu_{it}$$

under the assumption $y_{i0} = 0$.

4. The first 100 observation of y_{it} are dismissed to reduce the influence of the choice of the initial observations on the outcome.

Finally, parameters of the SVAR or SVECM model are estimated with two methods: ML and two steps quasi ML. In both estimation methods, four algorithms are used to search for the parameter values that maximize the likelihood function: three general maximization algorithms (BFGS, NEWTON and BHHH provided in the CML library in GAUSS) and the EM algorithm.

The outcomes, for each of the estimation methods and the maximization algorithms, are evaluated on the basis of:

- number of successful estimates (algorithm converges)
- ratio of estimates that satisfy the conditions (4) for $c = 0.01$
- mean and variance of the estimated parameters
- convergence to the true parameter values for increasing sample size
- sensitivity to choice of the initial values

4.2 Choice of parameters values

The Monte Carlo experiment was performed for three different lengths of the time series $T = 50, 150, 500$. Time dimensions $T = 50, 150$ correspond to lengths of time series used in the empirical analysis, whereas $T = 500$ captures the asymptotic behavior of examined estimators and maximization algorithms.

In both data generating processes, the residuals Bu_t were distributed according to the mixture of two normal distributions with the following parameter values:

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Psi = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \tag{5}$$

Two different proportion parameters were considered. Firstly, the mixture proportion was set to $\gamma = 0.5$, thus Bu_t was equally often distributed according to $N(0, BB')$ as to $N(0, B\Psi B')$. Finally $\gamma = 0.8$, which means that the second component was much more rarely observable. It was expected that the choice of γ would influence the small sample properties of the estimators in three ways: by effecting a rate of successful estimates, a frequency of choosing the true, rather then spurious, maximizers and efficiency (measured by estimator variance).

4.2.1 Structural Vector Autoregressive Model (SVAR)

In the first part of the experiment data was generated according to the VAR model with the order of autoregression $p = 1$.

$$y_t = A_0 + A_1 y_{t-1} + Bu_t \tag{6}$$

The autoregressive parameters were chosen to ensure that the process y_t was stationary

$$A_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, A = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \tag{7}$$

4.2.2 Structural Vector Error Correction Model (SVECM)

The order of autoregression is set as $p = 2$ and the model takes the following form

$$\Delta y_t = A_0 + \alpha\beta' y_{t-1} + \Gamma\Delta y_{t-1} + Bu_t \tag{8}$$

The parameters of the SVECM model were chosen to ensure that the process is well defined[1]

$$\begin{aligned} \alpha &= \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix}, \beta = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ A_0 &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Gamma = \begin{bmatrix} 0.2 & 0.5 \\ 0.5 & 0.2 \end{bmatrix} \end{aligned} \tag{9}$$

4.3 Results

Ratios of successful estimates Tables 1 and 2 present ratios of successful estimates for the VAR model, which are computed as the number of the outcomes with nonsingular covariance matrix and $0 < \gamma < 1$, divided by the total number of Monte Carlo iterations. Results indicate that the general maximization algorithms suffer many problems when estimating model parameters. More frequently, the parameters converge to singularity points or end up on the boundaries ($\gamma = 0$ or 1). This unwanted behavior is the strongest for the short time series ($T = 50$), when the ratios vary between 10% – 60% for algorithms that start with the true parameters values and 5% – 35% when they begin with false parameters values. For long time series ($T = 500$), the ratios are 70% – 90% and 30% – 80% respectively. In practice, we can expect the second case to occur more often and therefore, the results question the usage of this kind of algorithms. The EM algorithm outperforms the rest of algorithms in terms of the number of successful estimates. It converges to local maxima

[1] Let us denote by $C(z)$ the following polynomial

$$C(z) = (1 - z) I_k - \alpha\beta' z - \sum_{i=1}^{p-1} (1 - z) z\Gamma_i$$

Then, the VECM process is well defined if the following conditions hold

1. $\det(C(z)) = 0 \Rightarrow |z| \geq 1$
2. The number of unit roots $z = 1$, is exactly $k - r$, where $r = rk(\alpha) = rk(\beta)$

in almost all cases. Its disadvantage is, however, a very slow rate of convergence and lengthy time of computation (for more details see Redner and Walker (1984), McLachlan and Krishnan (1997)).

Tables 11 and 12 summarize the ratios of successful estimates[2] for the VECM model. For the two steps quasi ML method, they are qualitatively similar to those obtained in the VAR experiment. When the estimation procedures are initiated at true parameter values, the general maximization algorithms (NEWTON and BFGS) converge in $15-60\%$ cases for short time series $T=50$, compared with 90% for the EM algorithm. As the time dimension increases, differences between algorithms decrease and the ratios for general maximization algorithms reach almost 100%. When the estimation begins with parameter values that differ from the true ones, the ratios of successful estimates for the BFGS do not exceed 35% for all time lengths ($T=50,500$), whereas the NEWTON algorithm converges in $30-90\%$ cases depending on the time dimension. Both general maximization algorithms perform significantly worse then the EM algorithm, for which the rate of convergence is close to 100%.

When the ML method is considered, there appears to be more differences between the VAR and VECM experiments. The general maximization algorithms converge in around $20-30\%$ of the cases for $T=50$ and $80-100\%$ of the cases for $T=500$. The EM algorithm, however, does not perform significantly better and converges only in 40% of cases for $T=50$ and 95% of cases for $T=500$. These results indicate that the complexity of the estimation problem influences significantly the chances of successful convergence.

Finally, comparisons of different maximization algorithms bring two conclusions. Firstly, there are algorithms, such as BHHH[3], very sensitive to the length of the time series. For $T=50$, it falls far behind the BFGS and NEWTON algorithms. Secondly, BFGS is more frequently successful than the NEWTON algorithm when the initial guesses are close to the true parameters. The difference seems significant especially for very short time series. The results show, however, that the NEWTON algorithm is much more robust to the initial guesses of the parameters. Thirdly, the ratios of successful estimates and the true local maximizers hardly depend on the number of observations.

It is interesting to compare the results of ML and two steps quasi ML methods. It appears that the two steps quasi ML method leads more often to the successful estimates and to the true maximizers rather then the spurious ones. These preliminary results can not fully support the choice of this method in empirical applications, as the precision of estimates needs to be taken into account. However, it already indicates the advantages of simplifying the estimation problem.

[2] As in the VAR experiment the BHHH algorithm performes much worse then other algorithms, it is ommited in futher research.

[3] Comparison based on the VAR experiment

Autoregressive (VAR and VECM) parameters The comparison of the parameter estimates is based on the outcomes of the BFGS algorithm[4]. For all the two steps procedures, regardless of the maximization algorithm, the autoregressive parameters were estimated in the same way. Therefore, there is no need to compare results between the algorithms. Tables 5, 6, 15 and 16 present the means and the variances of the estimators for VAR and VECM models respectively. The outcomes satisfy condition (4) and are presented for the ML and two steps quasi ML separately. It is worth emphasizing that both methods produce very similar results. They confirm the consistency of the estimators, hence in all considered cases the mean converges to true parameter values and the variance decreases[5].

Mixture parameters Firstly, the estimates of the mixing parameters are compared on the basis of a ML with a BFGS maximization algorithm. Their properties (mean and the variance) are summarized in the Tables 7 and 17. The outcomes are less satisfying then in the autoregressive parameters case, but still show the consistency as the mean converges to the true parameter values and the variance decreases. It may be noticed that most of the problems arise while estimating the matrix Ψ. The biggest of the diagonal elements is estimated very imprecisely (its variance across Monte Carlo iterations reaches 313.04 for $T = 50$ for VAR and 331.26 for VECM model) and thereby influences the estimates of the rest of parameters.

Secondly, the results for three estimation procedures: a ML with BFGS (called M1), a two steps quasi ML with BFGS (called M2) and a two steps quasi ML with an EM (called EM2) are compared. The outcomes for the mixing proportion $\gamma = 0.5$ are illustrated in the Figures 1 and 2. It shows that the two steps quasi ML method with EM algorithm is the most precise in estimating the crucial Ψ matrix (when both the mean and the variance of the estimators are taken into account). For other mixture parameters, the outcomes are comparable across all three procedures (for more details see Tables 8-9 and Tables 18-19).

Spurious maximizers The importance of the spurious maximizers problem is illustrated by the results in Table 10. It summarizes the mean and the variance of the VAR and mixture parameters estimators for the cases in which the condition (4) is not satisfied. For $T = 50$, the mean of Ψ_2 estimators reaches almost 5000 and decreases to 2936 for $T = 150$. It means that in some cases the estimation procedures produce very unrealistic results which are characterized by high values of $\hat{\Psi}$ and low values of mixing proportion estimators (mean of $\hat{\gamma}$ was 0.193 and 0.117 for $T = 50, 150$ respectively).

[4] Results for other maximization algorithms are very simmilar and therefore they are not discussed in details.

[5] The t-ratios mean and variacne were also computed and they confirm good properties of the estimators (converge to the first two moments of N(0,1)). Tables that summarize the t-ratios are available upon request.

The autoregressive parameters estimators were not affected by the existence of the spurious maximizers. Even when the mixing parameters were estimated incorrectly, they were still similar to the results for cases in which (4) is satisfied and converged to true parameter values as the sample size increases. It suggests that the estimators of autoregressive parameters are robust to the choice of a local maximizer.

As previously discussed algorithms may converge to the spurious maximizers rather then to the true ones. To disregard these cases, the condition (4) was checked for every estimate. Tables 3, 4, 13 and 14 summarize the ratios of the number of true local maximizers to the number of successful estimates. The results show that the ratio increases with the length of the times series. For $T = 50$, it starts from 66% to 84%, whereas for $T = 500$ all the results exceed 99%. Unfortunately, the low ratio for short time series means that when the macroeconomic time series are used it may be expected that the spurious maximizers will arise quite often.

5 Conclusions

In this paper, we describe and discuss issues associated with an estimation of structural VAR models with mixtures of two normal distributions. The main theoretical difficulties that arise are a lack of global identifiability of parameters and an unbounded likelihood function. The first issue can be easily overcome because, under some mild restrictions, the parameters are locally identifiable and therefore, a ML estimation method can be applied. The second problem requires a new definition of a ML estimator because a global maximum of a likelihood function does not exist. Moreover, the likelihood function has many spurious local maxima, which make it difficult to find the proper ML estimates. We present how the issue is solved in the literature and adopt this approach to the SVAR models with a mixture distribution.

Finally, we perform a Monte Carlo experiment that compares different estimation methods and maximization algorithms. The outcomes indicate that there are no significant differences in the efficiency between the two discussed estimation methods: ML and two steps quasi ML. This result favours the two steps method as it is simpler and less computationally demanding. Next, we compare the properties of different maximization algorithms. The general maximization algorithms seem to perform worse then the EM algorithm. It is more frequent that they are not able to produce any results or lead to spurious maximizers. Estimates based on these methods vary more across the MC iterations, particularly for short time series. The differences between these two types of algorithms become negligible for long time series $T = 500$, when the ratio of successful estimations and the moments of the obtained estimators equalize. The main disadvantages of the EM algorithm are difficulties with computing the variance of the estimators[6] and the lengthy time of computations.

[6] To estimate asymptotic variance of the parameters some modification of the algorithm need to be introduced.

The experiment confirms that spurious maximizers are one of the crucial problems when estimating the parameters of SVAR models with the mixture of normal distributions. It happens that the estimates, which constitute local maxima of the likelihood function, are produced by a small group of observations with a low variance. Therefore, they give a high value of the likelihood function but do not represent a ML estimate with its statistical properties. The existence of spurious maximizers threatens the estimates of the mixing parameters but does not affect the estimates of the autoregressive parameters.

.

References

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrica 56*(3), 463–474.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistics Society Ser. B (methodological) 39*, 1–38.

Diebold, F. X. and G. D. Rudebusch (1996). Measuring business cycle: A modern perspective. *Review of Economics and Statistics 78*, 67–77.

Goodwin, T. H. (1993). Business-cycle analysis with a markov-switching model. *Journal of Business & Economic Statistics 11*(3), 331–339.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*, 357–384.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distribution. *The Annals of Statistics 18*(2), 795–800.

Kiefer, A. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist*, 887–906.

Kim, C.-J. and C. R. Nelson (1998). Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. *The Review of Economics and Statistics 80*(2), 188–201.

Kim, C.-J. and C. R. Nelson (1999). Has the U.S. economy become more stable? a bayesian approach based on a markow-switching model of the business cycle. *The Review of Economics and Statistics 81*(4), 608–616.

Krolzig, H.-M. (1997). *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Springer-Verlag, Berlin.

Lanne, M. and H. Lütkepohl (2005). Structural vector autoregressions with nonnormal residuals. CESinfo Working Paper No. 330.

Lanne, M. and H. Lütkepohl (2008). Identifying monetary policy shocks via changes in volatility. *Journal of Money, Credit and Banking 40*, 1131–1149.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer-Verlag, Berlin.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions.* Wiley.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models.* Wiley.

Redner, R. A. and H. F. Walker (1984). Mixture densities, maximum likelihood and the EM algorithm. *Society for Industrial and Appied Mathematics 26*, 195–239.

Rigobon, R. (2003). Identification through heteroscedasticity. *Review of Economics and Statistics 85*, 777–792.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica 39*(3), 577–591.

Sims, C. A. and T. Zha (2006). Were there regime switches in U.S. monetary policy? *American Economic Review 96*, 54–81.

Smith, A., P. A. Naik, and C.-L. Tsai (2006). Markov-switching model selection using kullback-leibler divergence. *Journal of Econometrics 134*, 553–577.

6 Appendix

6.1 "Label Switching"

We will show that for $\tilde{B} = B\Psi^{0.5}$, $\tilde{\Psi} = \Psi^{-1}$ and $\tilde{\gamma} = 1 - \gamma$ and any $\varepsilon \in R$ the following equality holds

$$f\left(\varepsilon; \tilde{B}, \tilde{\Psi}, \tilde{\gamma}\right) = f\left(\varepsilon; B, \Psi, \gamma\right)$$

The density function $f\left(\varepsilon; B, \Psi, \gamma\right)$ consists of two components

$$\begin{aligned} f\left(\varepsilon; B, \Psi, \gamma\right) &= \gamma \det\left(BB'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(BB'\right)^{-1} \varepsilon\right) \\ &\quad + \left(1-\gamma\right) \det\left(B\Psi B'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(B\Psi B'\right)^{-1} \varepsilon\right) \\ &= f_1\left(\varepsilon\right) + f_2\left(\varepsilon\right) \end{aligned}$$

Lets $\tilde{f}_1\left(\varepsilon\right)$ and $\tilde{f}_2\left(\varepsilon\right)$ denote the components of the density function computed for the new parameters vectors $\tilde{B}$, $\tilde{\Psi}$ and $\tilde{\gamma}$. Then the first component $\tilde{f}_1\left(\varepsilon\right) = f_2\left(\varepsilon\right)$

$$\begin{aligned} \tilde{f}_1\left(\varepsilon\right) &= \tilde{\gamma} \det\left(\tilde{B}\tilde{B}'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(\tilde{B}\tilde{B}'\right)^{-1} \varepsilon\right) \\ &= \left(1-\gamma\right) \det\left(B\Psi^{0.5}\Psi'^{0.5}B'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(B\Psi^{0.5}\Psi'^{0.5}B'\right)^{-1} \varepsilon\right) \\ &= \left(1-\gamma\right) \det\left(B\Psi B'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(B\Psi B'\right)^{-1} \varepsilon\right) \\ &= f_2\left(\varepsilon\right) \end{aligned}$$

and the second one $\tilde{f}_2\left(\varepsilon\right) = f_1\left(\varepsilon\right)$

$$\begin{aligned} \tilde{f}_2\left(\varepsilon\right) &= \left(1-\tilde{\gamma}\right) \det\left(\tilde{B}\tilde{\Psi}\tilde{B}'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(\tilde{B}\tilde{\Psi}\tilde{B}'\right)^{-1} \varepsilon\right) \\ &= \gamma \det\left(B\Psi^{0.5}\Psi^{-1}\Psi'^{0.5}B'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(B\Psi^{0.5}\Psi^{-1}\Psi'^{0.5}B'\right)^{-1} \varepsilon\right) \\ &= \gamma \det\left(BB'\right)^{-0.5} \exp\left(-0.5\varepsilon' \left(BB'\right)^{-1} \varepsilon\right) \\ &= f_1\left(\varepsilon\right) \end{aligned}$$

Finally,

$$\begin{aligned} f\left(\varepsilon; \tilde{B}, \tilde{\Psi}, \tilde{\gamma}\right) &= \tilde{f}_1\left(\varepsilon\right) + \tilde{f}_2\left(\varepsilon\right) \\ &= f_2\left(\varepsilon\right) + f_1\left(\varepsilon\right) \\ &= f\left(\varepsilon; B, \Psi, \gamma\right) \end{aligned}$$

7 Results: SVAR

Table 1: VAR. Ratio of successful estimates, algorithms initiated with the true parameters values.

		ML				two steps quasi ML			
γ	T	BFGS	NEWTON	BHHH	EM	BFGS	NEWTON	BHHH	EM
0.5	50	0.592	0.262	0.102	1.000	0.625	0.334	0.17	1.00
	150	0.896	0.515	0.611	0.996	0.882	0.498	0.583	0.994
	500	0.995	0.734	0.972	0.992	0.992	0.735	0.969	0.992
0.8	50	0.384	0.165	0.016	0.998	0.410	0.186	0.023	1.00
	150	0.758	0.403	0.230	0.993	0.733	0.415	0.228	0.992
	500	0.979	0.635	0.749	0.990	0.976	0.647	0.757	0.919

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 2: VAR. Ratio of successful estimates, algorithms not initiated with the true parameters values.

		ML				two steps quasi ML			
γ	T	BFGS	NEWTON	BHHH	EM	BFGS	NEWTON	BHHH	EM
0.5	50	0.212	0.287	0.060	1.000	0.242	0.344	0.123	0.999
	500	0.306	0.727	0.768	0.992	0.275	0.728	0.660	0.989
0.8	50	0.161	0.272	0.058	0.998	0.160	0.371	0.046	0.998
	500	0.584	0.822	0.628	0.994	0.385	0.820		0.990

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 3: VAR. Ratio of successful estimates that satisfy condition (4) for $c = 0.01$ to all successful estimates, algorithms initiated with the true parameters values.

		ML				two steps quasi ML			
γ	T	BFGS	NEWTON	BHHH	EM	BFGS	NEWTON	BHHH	EM
0.5	50	0.775	0.786	0.863	0.812	0.913	0.904	0.935	0.946
	150	0.948	0.940	0.957	0.948	0.984	0.970	0.992	0.989
	500	0.998	0.997	0.998	0.998	0.999	0.999	0.998	1.00
0.8	50	0.930	0.915	0.875	0.903	0.971	0.962	0.956	0.919
	150	0.991	0.985	1.00	0.75	0.999	0.995	1.00	0.986
	500	1.00	1.00	1.00	0.999	1.00	1.00	1.00	1.00

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 4: VAR. Ratio of successful estimates that satisfy condition (4) for $c = 0.01$ to all successful estimates, algorithms not initiated with the true parameters values.

		ML				two steps quasi ML			
γ	T	BFGS	NEWTON	BHHH	EM	BFGS	NEWTON	BHHH	EM
0.5	50	0.901	0.840	0.800	0.749	0.967	0.936	1.00	0.930
	500	0.997	1.00	0.996	0.999	0.996	1.00	0.997	0.999
0.8	50	0.969	0.893	0.810	0.850	0.962	0.921	0.956	0.944
	500	1.00	0.998	0.995	0.999	1.00	0.999		1.00

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (6) and (7). T and γ denote the length of the sample and a mixing proportion parameter, respectively.

Table 5: VAR. The mean and the variance of the autoregressive parameters estimates for the two steps quasi ML method (BFGS algorithm iniciated with the true parameters values).

Parameters		$A_1^{(0)}$	$A_2^{(0)}$	A_{11}	A_{21}	A_{12}	A_{22}
True values		0	0	0.5	0	0	0.5
γ	T	Mean					
0.5	50	-0.0009	0.0090	0.4385	0.0066	-0.0056	0.4440
	150	-0.0051	-0.0035	0.4794	-0.0015	-0.0022	0.4803
	500	-0.0014	-0.0010	0.4937	-0.0029	-0.0007	0.4940
0.8	50	-0.0058	-0.0085	0.4434	-0.0122	-0.0026	0.4510
	150	-0.0006	0.0009	0.4774	-0.0023	-0.0001	0.4782
	500	0.0006	0.0021	0.4921	-0.0006	-0.0002	0.4935
		Variance					
0.5	50	0.0247	0.0855	0.0152	0.0529	0.0055	0.0156
	150	0.0076	0.0215	0.0046	0.0168	0.0018	0.0049
	500	0.0021	0.0059	0.0015	0.0045	0.0005	0.0016
0.8	50	0.0269	0.0481	0.0171	0.0366	0.0099	0.0129
	150	0.0071	0.0132	0.0053	0.0102	0.0029	0.0052
	500	0.0018	0.0039	0.0015	0.0027	0.0008	0.0016

NOTE: The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 6: VAR. The mean and the variance of the autoregressive parameters estimates for the ML method (BFGS algorithm iniciated with the true parameters values).

Parameters		$A_1^{(0)}$	$A_2^{(0)}$	A_{11}	A_{21}	A_{12}	A_{22}
True values		0	0	0.5	0	0	0.5
γ	T	Mean					
0.5	50	0.0025	-0.0056	0.4492	-0.0176	-0.0054	0.4535
	150	-0.0062	-0.0010	0.4806	0.0000	-0.0030	0.4851
	500	-0.0014	-0.0007	0.4936	-0.0043	-0.0006	0.4953
0.8	50	0.0030	-0.0080	0.4429	-0.0196	-0.0069	0.4613
	150	0.0000	0.0003	0.4783	0.0010	-0.0003	0.4832
	500	0.0008	0.0015	0.4919	-0.0010	-0.0001	0.4947
		Variance					
0.5	50	0.0285	0.099	0.0175	0.0602	0.0061	0.0190
	150	0.0082	0.0186	0.0048	0.0157	0.0018	0.0049
	500	0.0021	0.0051	0.0016	0.0038	0.0005	0.0014
0.8	50	0.0358	0.0448	0.0184	0.0410	0.0110	0.0164
	150	0.0074	0.0114	0.0053	0.0084	0.0030	0.0040
	500	0.0019	0.0032	0.0015	0.0021	0.0009	0.0013

NOTE: The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 7: VAR. The mean and the variance of the mixing parameter estimates for the ML method (BFGS algorithm iniciated with the true parameters values).

Parameters		B_{11}	B_{21}	B_{12}	B_{22}	Ψ_1	Ψ_2	γ
True values		1	0	0	1	1	5	0.5/0.8
γ	T	Mean						
0.5	50	0.9696	0.0041	-0.0008	0.6739	1.1309	18.541	0.5183
	150	1.0072	-0.0006	0.0000	0.8503	1.0355	10.008	0.4939
	500	1.0054	-0.0011	-0.0005	0.9607	0.9791	6.0943	0.5031
0.8	50	0.9413	-0.0131	-0.0071	0.6328	1.2300	15.284	0.6339
	150	0.9816	-0.0222	0.0090	0.8626	0.9604	7.2953	0.7134
	500	0.9958	-0.0010	0.0016	0.9635	0.9544	5.4723	0.7717
		Variance						
0.5	50	0.0716	0.3132	0.0306	0.0862	4.2206	340.36	0.0353
	150	0.0284	0.1302	0.0147	0.0711	0.4660	120.44	0.0356
	500	0.0008	0.0388	0.0037	0.0314	0.1051	12.934	0.0181
0.8	50	0.0540	0.1640	0.0337	0.0533	6.7184	223.14	0.0346
	150	0.0153	0.0679	0.0187	0.0343	0.4133	19.281	0.0299
	500	0.0030	0.0181	0.0067	0.0088	0.1301	1.7017	0.0110

NOTE: The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 8: VAR. The mean and the variance of the mixing parameter estimates for the two steps quasi ML method (BFGS algorithm iniciated with the true parameters values).

Parameters		B_{11}	B_{21}	B_{12}	B_{22}	Ψ_1	Ψ_2	γ
True values		1	0	0	1	1	5	0.5/0.8
γ	T	Mean						
0.5	50	0.9791	0.0151	−0.0016	0.7006	1.0384	16.008	0.4562
	150	1.0021	0.0064	−0.003	0.8605	0.9686	9.4433	0.4798
	500	1.0047	−0.0001	−0.0008	0.9739	0.9808	5.8277	0.5047
0.8	50	0.9435	0.0127	−0.0101	0.6837	0.9831	11.008	0.6201
	150	0.9778	−0.0161	0.0080	0.8734	0.9754	6.8268	0.7101
	500	0.9954	−0.0007	0.0012	0.9689	0.9565	5.3703	0.7727
		Variance						
0.5	50	0.0688	0.3102	0.0277	0.0972	7.5682	313.04	0.0488
	150	0.0312	0.1436	0.0171	0.0759	0.4699	111.72	0.0393
	500	0.0078	0.0313	0.0039	0.0300	0.1029	9.9939	0.0181
0.8	50	0.0382	0.1795	0.0361	0.0548	1.8765	83.093	0.0417
	150	0.0153	0.0656	0.0192	0.0344	0.5007	16.762	0.0311
	500	0.0029	0.0183	0.0069	0.0083	0.1258	1.5437	0.0107

NOTE: The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 9: VAR. The mean and the variance of the mixing parameter estimates for the two steps quasi ML method (EM algorithm iniciated with the true parameters values).

Parameters		B_{11}	B_{21}	B_{12}	B_{22}	Ψ_1	Ψ_2	γ
True values		1	0	0	1	1	5	0.5/0.8
γ	T	Mean						
0.5	50	0.943	0.0260	−0.0045	0.8453	0.8848	11.443	0.5333
	500	1.0049	0.0020	−0.0011	0.9765	0.9780	5.7325	0.5046
0.8	50	0.9669	−0.0258	0.0124	0.8504	0.7101	8.3221	0.7634
	500	0.9968	0.0003	0.0011	0.9735	0.9471	5.3650	0.7780
		Variance						
0.5	50	0.0582	0.3278	0.0334	0.1064	6.8999	189.44	0.0467
	500	0.0074	0.0304	0.0039	0.0275	0.0964	8.2750	0.0163
0.8	50	0.0280	0.1730	0.0546	0.0495	0.4517	65.635	0.0395
	500	0.0029	0.0173	0.0067	0.0077	0.1339	1.4941	0.0094

NOTE: The data generating process is described by (5), (6) and (7). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 10: VAR, The mean and the variance of estimators for the ML method, the mixing proportion $\gamma = 0.5$. The data generating process is described by (6), (7) and (5).

T	50		150	
	Mean	Var	Mean	Var
$A_1^{(0)}$	-0.0275	0.0367	0.0046	0.0097
$A_2^{(0)}$	-0.0689	0.1363	0.0269	0.0406
$A_{1,1}$	0.4370	0.0180	0.4845	0.0064
$A_{2,1}$	0.01741	0.0957	-0.0348	0.0281
$A_{1,2}$	-0.0034	0.0073	0.0072	0.0020
$A_{2,2}$	0.4480	0.0294	0.4817	0.0061
$B_{1,1}$	1.1507	0.2199	1.0997	0.1557
$B_{2,1}$	0.0588	0.3988	-0.0215	0.2382
$B_{1,2}$	-0.0008	0.0008	0.0036	0.0009
$B_{2,2}$	0.0724	0.0027	0.0869	0.0028
Ψ_1	40.557	77630	40.050	60582
Ψ_2	4988.49	$1.77e + 008$	2936.21	55389027
γ	0.1929	0.0032	0.1168	0.0013

8 Results: SVECM

Table 11: VECM. Ratio of successful estimates, algorithms initiated with the true parameters values.

		ML			two steps quasi ML		
γ	T	BFGS	NEWTON	EM	BFGS	NEWTON	EM
0.5	50	0.366	0.218	0.403	0.627	0.336	0.976
	150	0.882	0.588	0.801	0.919	0.566	0.996
	500	0.987	0.843	0.970	0.993	0.750	0.988
0.8	50	0.245	0.174	0.350	0.343	0.166	0.949
	150	0.726	0.501	0.706	0.740	0.381	0.987
	500	0.969	0.785	0.935	0.975	0.667	0.993

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 12: VECM. Ratio of successful estimates, algorithms not initiated with the true parameters values.

		ML			two steps quasi ML		
γ	T	BFGS	NEWTON	EM	BFGS	NEWTON	EM
0.5	50	0.130	0.144	0.268	0.241	0.336	0.999
	500	0.349	0.716	0.891	0.293	0.739	0.987
0.8	50	0.112	0.140	0.299	0.172	0.366	1.000
	500	0.287	0.759	0.931	0.329	0.847	0.988

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 13: VECM. Ratio of successful estimates that satisfy condition (4) for $c = 0.01$ to all sucesfull estimates, algorithms initiated with the true parameters values.

		ML			two steps quasi ML		
γ	T	BFGS	NEWTON	EM	BFGS	NEWTON	EM
0.5	50	0.839	0.972	0.990	0.907	0.881	0.944
	150	0.926	0.995	0.999	0.979	1	0.987
	500	0.998	1	1	0.999	0.999	0.999
0.8	50	0.894	0.977	0.991	0.983	0.952	0.969
	150	0.983	1	1	0.996	1	0.985
	500	1	1	1	1	1	0.999

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 14: VECM. Ratio of successful estimates that satisfy condition (4) for $c = 0.01$ to all sucesfull estimates, algorithms not initiated with the true parameters values.

		ML			two steps quasi ML		
γ	T	BFGS	NEWTON	EM	BFGS	NEWTON	EM
0.5	50	0.854	0.951	1	0.975	0.881	0.913
	500	0.997	1	1	0.997	0.999	0.999
0.8	50	0.866	0.971	1	0.994	0.937	0.950
	500	1	1	1	1	0.999	0.999

NOTE: Two methods are considered: Maximum Likelihood and two steps quasi Maximum Likelihood. For each estimation method four maximization algorithms are evaluated: BFGS, NEWTON, BHHH and EM. The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 15: VECM. The mean and the variance of the parameters estimates for the two steps quasi ML method (BFGS algorithm initiated with the true parameters values).

Parameters		β_2	α_1	α_2	$A_1^{(0)}$	$A_2^{(0)}$	Γ_{11}	Γ_{21}	Γ_{12}	Γ_{22}
True values		-1	-0.1	0.1	0	0	0.2	0.5	0.5	0.2
γ	T	Mean								
0.5	50	-1.628	-0.172	0.174	0.105	0.087	0.177	0.429	0.428	0.216
	150	-1.004	-0.125	0.132	-0.012	0.019	0.191	0.476	0.472	0.210
	500	-1.000	-0.108	0.111	-0.002	-0.001	0.197	0.490	0.491	0.205
0.8	50	-1.287	-0.180	0.157	-0.042	-0.062	0.178	0.426	0.427	0.190
	150	-1.007	-0.124	0.127	-0.017	-0.004	0.192	0.469	0.476	0.202
	500	-1.001	-0.107	0.110	-0.002	0.000	0.197	0.491	0.493	0.204
		Variance								
0.5	50	225.420	0.015	0.040	2.471	5.720	0.011	0.031	0.015	0.039
	150	0.039	0.002	0.006	0.125	0.242	0.003	0.009	0.003	0.010
	500	0.000	0.000	0.001	0.012	0.019	0.001	0.002	0.001	0.003
0.8	50	1595.93	0.016	0.027	1.575	3.309	0.013	0.024	0.017	0.029
	150	0.036	0.003	0.005	0.125	0.206	0.004	0.007	0.004	0.008
	500	0.001	0.001	0.001	0.010	0.013	0.001	0.002	0.001	0.002

NOTE: The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 16: VECM. The mean and the variance of the parameters estimates for the ML method (BFGS algorithm initiated with the true parameters values).

Parameters		β_2	α_1	α_2	$A_1^{(0)}$	$A_2^{(0)}$	Γ_{11}	Γ_{21}	Γ_{12}	Γ_{22}
True values		-1	-0.1	0.1	0	0	0.2	0.5	0.5	0.2
γ	T	Mean								
0.5	50	-0.999	-0.193	0.180	0.025	0.179	0.182	0.441	0.407	0.222
	150	-1.002	-0.125	0.126	0.000	0.006	0.192	0.480	0.472	0.208
	500	-1.000	-0.108	0.110	0.000	-0.003	0.197	0.492	0.491	0.205
0.8	50	-0.979	-0.196	0.162	0.116	-0.116	0.182	0.453	0.414	0.207
	150	-0.999	-0.125	0.123	-0.011	-0.019	0.196	0.478	0.474	0.203
	500	-1.001	-0.107	0.108	0.000	-0.001	0.197	0.493	0.493	0.203
		Variance								
0.5	50	0.045	0.016	0.035	1.653	2.048	0.014	0.036	0.016	0.040
	150	0.009	0.002	0.006	0.128	0.180	0.003	0.008	0.003	0.009
	500	0.000	0.000	0.001	0.012	0.018	0.001	0.002	0.001	0.002
0.8	50	0.060	0.015	0.021	1.064	0.999	0.017	0.029	0.017	0.029
	150	0.010	0.003	0.004	0.109	0.118	0.004	0.006	0.004	0.007
	500	0.000	0.001	0.001	0.009	0.011	0.001	0.002	0.001	0.002

NOTE: The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 17: VECM. The mean and the variance of the mixing parameters estimates for the ML method (BFGS algorithm initiated with the true parameters values).

Parameters		B_{11}	B_{21}	B_{12}	B_{22}	Ψ_1	Ψ_2	γ
True values		1	0	0	1	1	5	0.5/0.8
γ	T	Mean						
0.5	50	0.953	-0.018	0.011	0.648	1.059	19.615	0.551
	150	0.987	-0.006	-0.001	0.828	1.021	11.012	0.494
	500	0.997	-0.009	0.001	0.953	0.990	6.183	0.499
0.8	50	0.902	0.078	-0.018	0.612	1.202	16.464	0.649
	150	0.979	-0.002	0.000	0.849	0.949	7.593	0.710
	500	0.994	0.001	0.000	0.964	0.945	5.432	0.773
		Variance						
0.5	50	0.071	0.342	0.032	0.076	2.599	331.26	0.029
	150	0.031	0.144	0.016	0.079	1.524	169.026	0.036
	500	0.008	0.034	0.004	0.033	0.107	15.931	0.019
0.8	50	0.053	0.172	0.026	0.052	2.768	225.49	0.031
	150	0.016	0.073	0.020	0.036	0.781	39.406	0.033
	500	0.003	0.019	0.007	0.009	0.115	1.576	0.011

NOTE: The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 18: VECM. The mean and the variance of the mixing parameters estimates for the two steps quasi ML method (BFGS algorithm initiated with the true parameters values).

Parameters		B_{11}	B_{21}	B_{12}	B_{22}	Ψ_1	Ψ_2	γ
True values		1	0	0	1	1	5	0.5/0.8
γ	T	Mean						
0.5	50	0.945	-0.003	-0.002	0.680	1.192	14.223	0.442
	150	0.985	-0.017	-0.001	0.856	1.097	9.458	0.475
	500	0.997	-0.012	0.002	0.970	0.992	5.750	0.500
0.8	50	0.936	0.008	-0.008	0.671	0.966	11.231	0.593
	150	0.973	0.002	-0.002	0.876	0.930	6.969	0.711
	500	0.994	-0.002	0.001	0.971	0.946	5.328	0.773
		Variance						
0.5	50	0.078	0.317	0.030	0.091	0.924	241.47	0.048
	150	0.032	0.151	0.018	0.082	0.638	122.55	0.044
	500	0.008	0.033	0.004	0.030	0.102	6.488	0.018
0.8	50	0.043	0.161	0.029	0.054	1.539	162.63	0.045
	150	0.014	0.072	0.021	0.037	0.344	32.394	0.037
	500	0.003	0.019	0.007	0.008	0.114	4.015	0.011

NOTE: The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

Table 19: VECM. The mean and the variance of the mixing parameters estimates for the two steps quasi ML method (EM algorithm initiated with the true parameters values).

Parameters		B_{11}	B_{21}	B_{12}	B_{22}	Ψ_1	Ψ_2	γ
True values		1	0	0	1	1	5	0.5/0.8
γ	T	Mean						
0.5	50	0.957	-0.001	-0.003	0.839	0.828	10.781	0.531
	150	0.989	-0.015	0.003	0.898	0.940	8.398	0.496
	500	0.997	-0.014	0.003	0.972	0.991	5.688	0.499
0.8	50	0.943	-0.009	0.004	0.864	0.767	7.235	0.787
	150	0.980	-0.004	0.002	0.905	0.788	6.782	0.739
	500	0.994	-0.002	0.002	0.973	0.939	5.266	0.775
		Variance						
0.5	50	0.059	0.352	0.039	0.103	0.739	178.21	0.050
	150	0.029	0.152	0.021	0.077	0.435	90.99	0.040
	500	0.008	0.033	0.005	0.027	0.101	6.139	0.016
0.8	50	0.027	0.149	0.052	0.046	0.685	55.824	0.037
	150	0.013	0.081	0.028	0.036	0.342	29.583	0.037
	500	0.003	0.020	0.007	0.008	0.115	1.538	0.010

NOTE: The data generating process is described by (5), (8) and (9). We denote by T and γ the length of the sample and a mixing proportion parameter, respectively.

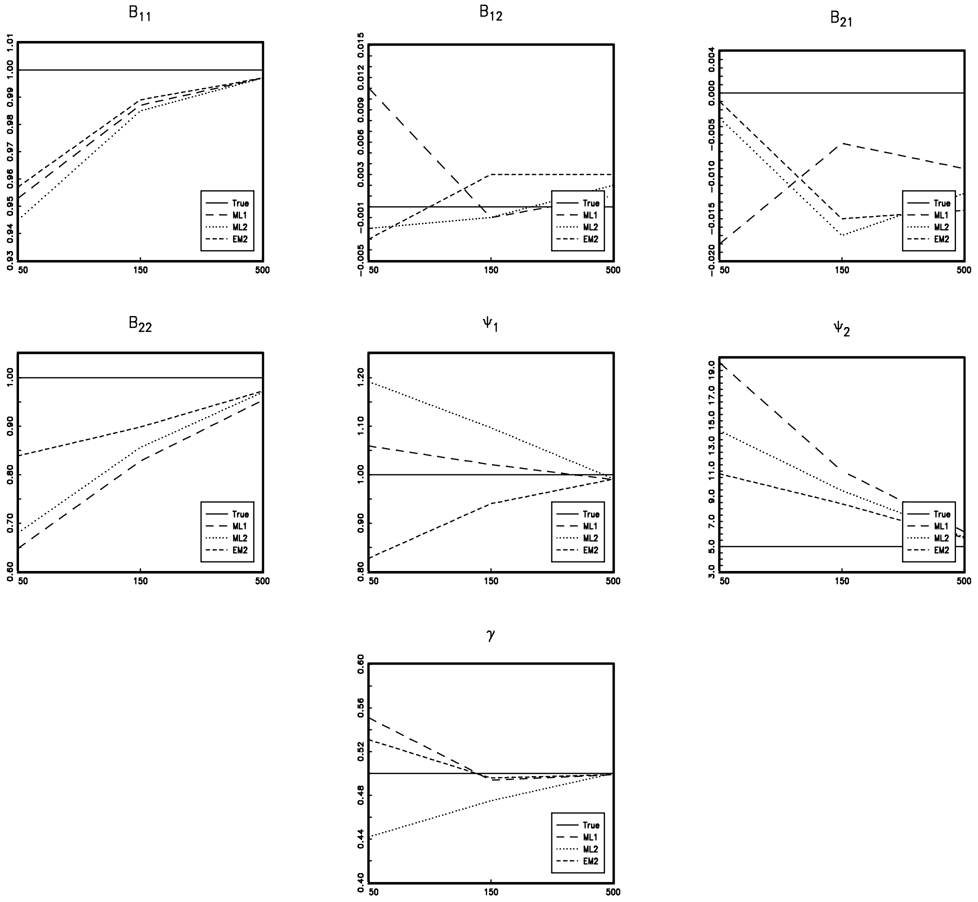

Figure 1: The mean of the estimates of mixture parameters for VECM conditional on the sample length. "True" describes the true parameter values whereas ML1, ML2 and EM2 present the results for the ML method with BFGS algorithm, two steps quasi ML method with BFGS algorithm and two steps quasi ML method with EM algorithm, respectively. The data generating process is described by (5), (8) and (9).

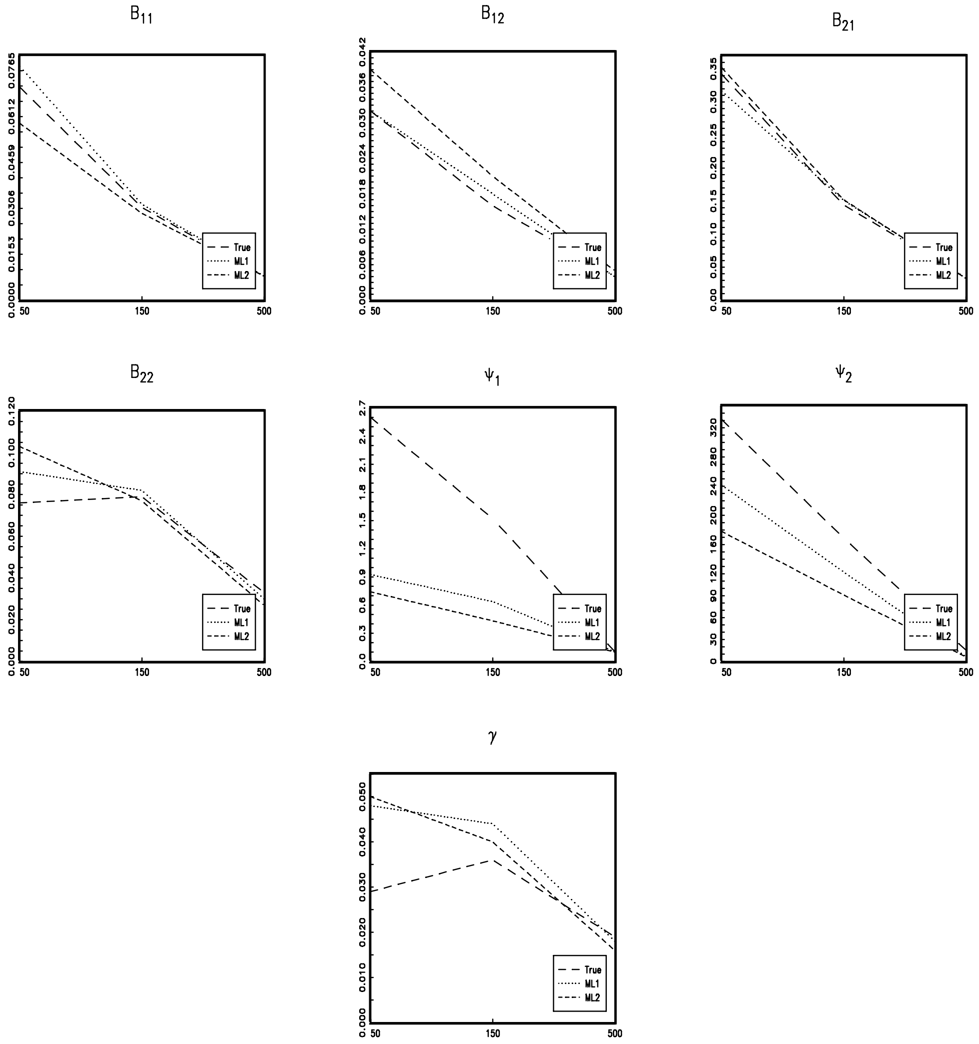

Figure 2: The variance of the estimates of mixture parameters for VECM conditional on the sample length. ML1, ML2 and EM2 present the results for the ML method with BFGS algorithm, two steps quasi ML method with BFGS algorithm and two steps quasi ML method with EM algorithm, respectively. The data generating process is described by (5), (8) and (9).

European
University
Institute

European University Press