A Comparison of Estimation Methods for Vector Autoregressive Moving-Average Models*

Christian Kascha†

Norges Bank

April 20, 2010

Abstract

Recently, there has been a renewed interest in modeling economic time series by vector autoregressive moving-average models. However, this class of models has been unpopular in practice because of estimation problems and the complexity of the identification stage. These disadvantages could have led to the dominant use of vector autoregressive models in macroeconomic research. In this paper, several simple estimation methods for vector autoregressive moving-average models are compared among each other and with pure vector autoregressive modeling using ordinary least squares by means of a Monte Carlo study. Different evaluation criteria are used to judge the relative performances of the algorithms.

JEL classification: C32, C15, C63
Keywords: VARMA Models, Estimation Algorithms, Forecasting

*I would like to thank Helmut Lütkepohl and Anindya Banerjee for helpful comments and discussion. The research for this paper has been done at the European University Institute, Florence. The views expressed in this paper are my own and do not necessarily reflect the views of Norges Bank.

†Christian Kascha, Norges Bank, Research Department, Bankplassen 2, 0107 Oslo, Norway. Telephone: +47 22 31 67 19. Fax: +47 22 42 40 62. christian.kascha@norges-bank.no

1 Introduction

Although vector autoregressive moving-average (VARMA) models have theoretical advantages compared to simpler vector autoregressive (VAR) models, VARMA models are rarely used in applied macroeconomic work. The likely reasons are estimation problems and, in particular, the complexity of the identification stage. This paper investigates the relative performance of several simple estimation methods for VARMA models that have not been compared systematically by means of a Monte Carlo study. The methods are also compared with maximum likelihood estimation and pure vector autoregressive modeling using ordinary least squares. The evaluation criteria are the accuracy of the parameter estimates, the accuracy of point forecasts as well as the precision of the estimated impulse responses. I focus on sample lengths and processes that could be considered typical for macroeconomic applications.

The problem of estimating VARMA models received considerable attention for several reasons. Linear models such as VARs or univariate autoregressive moving-average (ARMA) models have proved to be simple and analytically tractable, while capable of reproducing complex dynamics. Linear forecasts often appear to be more robust than nonlinear alternatives and their empirical usefulness has been documented in various studies (e.g. Newbold & Granger 1974). A more recent example is the integrated moving-average model for US inflation of Stock & Watson (2007). Therefore, VARMA models are of interest as generalizations of successful univariate ARMA models.

In the class of multivariate linear models, pure VARs dominate in macroeconomic applications. However, VAR models may require a rather large lag length in order to describe a series "adequately". This means a loss of precision because many parameters have to be estimated. The problem could be avoided by using VARMA models that may provide a more parsimonious description of the data generating process (DGP). In contrast to the class of VARMA models, the class of VAR models is not closed under linear transformations. For example, a subset of variables generated by a VAR process is typically generated by a VARMA, not by a VAR process (Lütkepohl 1984*a*,*b*). The VARMA class includes many models of interest such as unobserved component models. It is well known that linearized dynamic stochastic general

equilibrium (DSGE) models imply that the variables of interest are generated by a finite-order VARMA process. Fernández-Villaverde, Rubio-Ramírez, Sargent & Watson (2007) show formally how DSGE models and VARMA processes are linked. Also Cooley & Dwyer (1998) claim that modeling macroeconomic time series systematically as pure VARs is not justified by the underlying economic theory. The recent debate between Chari, Kehoe & McGrattan (2008) and Christiano, Eichenbaum & Vigfusson (2006) on the ability of structural VARs to uncover fundamental shocks also questions implicitly the ability of pure VARs to capture the dynamics of an economic system.

However, there are also some complications that make VARMA modeling more difficult. First, VARMA representations are not unique. That is, there are typically many parameterizations that can describe the same DGP (see Lütkepohl 2005). Therefore, a researcher has to choose first an identified representation. In any case, an identified VARMA representation has to be specified by more integer-valued parameters than a VAR representation that is determined just by one integer parameter, the lag length. This aspect introduces additional uncertainty at the specification stage of the modeling process, although procedures for VARMA models do exist which could be used in a completely automatic way (Hannan & Kavalieris 1984*b*, Poskitt 1992). An identified representation, however, is needed for consistent estimation. Apart from a more involved specification stage, the estimation stage is also affected by the identification problem because one usually has to examine many different models which turn out not to be identified ex-post.

The literature on the estimation of VARMA models traditionally focussed on maximum likelihood methods which are asymptotically efficient (e.g. Hillmer & Tiao 1979, Mauricio 1995, Metaxoglou & Smith 2007). However, several simpler estimation methods have also been proposed as computationally less intense and more robust alternatives to maximum likelihood (see e.g. Durbin 1960, Hannan & Rissanen 1982, Hannan & Kavalieris 1984*b*, Koreisha & Pukkila 1990, Kapetanios 2003, Dufour & Pelletier 2008). In addition, they can serve to initialize maximum likelihood procedures and they can be used in a foregoing specification search.[1] However, it is not clear which of these methods is preferable under

[1]Recently, subspace algorithms for state space systems, an equivalent representation of a VARMA process, have become popular also among econometricians. Examples are the algorithms of Van Overschee & DeMoor

which circumstances. The available comparisons in the literature are relatively limited. The above cited papers include comparisons of some of the simple estimation algorithms but either consider only a limited number of algorithms or only one or two VARMA DGPs. Kapetanios (2003) considers a wide range of algorithms and DGPs but only considers one low-dimensional and relatively well behaved VARMA process.

In contrast, this study focusses on simple algorithms and compares them using many different DGPs. The eigenvalues of both the autoregressive and the moving-average polynomial of a given VARMA process are varied in order to investigate the algorithms' performance in favorable and difficult cases. This is important as the algorithms have to work well in a variety of situations because the underlying DGP is unknown in applications. Also, instead of focussing only on the accuracy of the parameter estimates, I consider the use of the estimated VARMA models. After all, a researcher might be rather interested in the accuracy of the generated forecasts or the precision of the estimated impulse response function than in the actual parameter estimates. To the best of my knowledge, this is the only study on VARMA estimation that shares these features. I conduct Monte Carlo simulations for four different DGPs with varying parameterizations. I consider the case when the orders of the true process are known and I focus on stationary processes. Four different simple algorithms are used and compared among each other and with two benchmark VARs. They are benchmarked against a full information maximum likelihood procedure starting from the true parameter values. The algorithms are the Hannan-Rissanen procedure (Durbin 1960, Hannan & Rissanen 1982), the iterative least squares procedure of Kapetanios (2003), the generalized least squares procedure of Koreisha & Pukkila (1990) and the Hannan-Kavalieris algorithm (Hannan & Kavalieris 1984*b*). The obtained results suggest that the algorithm of Hannan & Kavalieris (1984*b*) is generally preferable to the other algorithms and the benchmark VARs. However, the procedure is technically not very reliable in that the algorithm very often yields estimated models which are not stable or produces too many outliers for specific DGPs and parameterizations. Therefore, the algorithm would have to be improved in order to make it an alternative tool for applied researchers.

(1994) or Larimore (1983). See also the survey of Bauer (2005). A comparison with these estimators is however beyond the scope of the paper.

The rest of the paper is organized as follows. In section 2 stationary VARMA processes are introduced and the Echelon parameterization is presented. In section 3 the different estimation algorithms are described. The setup and the results of the Monte Carlo study are presented in section 4. Section 5 concludes. All programs are written in GAUSS and can be obtained from the homepage of the author.

2 Stationary VARMA Processes

I consider linear, time-invariant, covariance - stationary processes $(y_t)_{t\in\mathbb{Z}}$ of dimension K that allow for a VARMA(p,q) representation of the form

$$A_0 y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + M_0 u_t + M_1 u_{t-1} + \ldots + M_q u_{t-q} \quad (1)$$

for $t \in \mathbb{Z},\ p, q \in \mathbb{N}_0$. The matrices $A_0, A_1, \ldots, A_p$ and $M_0, M_1, \ldots, M_q$ are of dimension $(K \times K)$. The term u_t represents a white noise sequence of random variables with mean zero and nonsingular covariance matrix Σ. In principle, equation (1) should contain an intercept term and other deterministic terms in order to account for random series with non-zero mean and/or seasonal patterns. This has not been done here in order to simplify the exposition of the basic properties of VARMA models and the related estimation algorithms. For most of the algorithms discussed later, it is assumed that the mean has been subtracted prior to estimation. We consider models of the form (1) such that $A_0 = M_0$ and $A_0,\ M_0$ are nonsingular. This does not imply a loss of generality as long as no variable can be written as a linear combination of the other variables (Lütkepohl 2005). It can be shown that any stationary and invertible VARMA process can then be expressed in the above form.

Let L denote the lag-operator, i.e. $Ly_t = y_{t-1}$ for all $t \in \mathbb{Z}$, $A(L) = A_0 - A_1 L - \ldots - A_p L^p$ and $M(L) = M_0 + M_1 L + \ldots + M_q L^q$. We can write (1) more compactly as

$$A(L) y_t = M(L) u_t,\ t \in \mathbb{Z}. \quad (2)$$

VARMA processes are stationary and invertible if the roots of these polynomials are all outside

the unit circle. That is, if

$$|A(z)| \neq 0,\ |M(z)| \neq 0 \text{ for } z \in \mathbb{C}, |z| \leq 1$$

is true, where $|\cdot|$ denotes the determinant of a matrix. These restrictions are important for the estimation and for the interpretation of VARMA models. The first condition ensures that the process is covariance-stationary and has an infinite moving-average or canonical moving-average representation

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} = \Phi(L) u_t, \tag{3}$$

where $\Phi(L) = A(L)^{-1} M(L)$. If $A_0 = M_0$ is assumed, then $\Phi_0 = I_K$ where I_K denotes an identity matrix of dimension $(K \times K)$. The second condition ensures the invertibility of the process, in particular the existence of an infinite autoregressive representation

$$y_t = \sum_{i=1}^{\infty} \Pi_i y_{t-i} + u_t, \tag{4}$$

where $A_0 = M_0$ is assumed and $\Pi(L) = I_K - \sum_{i=1}^{\infty} \Pi_i L^i = M(L)^{-1} A(L)$. This representation indicates, why a pure VAR with a large lag length might well approximate mixed VARMA processes.

It is well known that the representation in (1) is generally not identified unless special restrictions are imposed on the coefficient matrices (Lütkepohl 2005). Precisely, all pairs of polynomial matrices $(A(L), M(L))$ which lead to the same canonical moving-average operator $\Phi(L) = A(L)^{-1} M(L)$ are equivalent. However, uniqueness of the pair $(A(L), M(L))$ is required for consistent estimation. The *Echelon* representation is based on the Kronecker index theory introduced by Akaike (1974). A VARMA representation for a K-dimensional series y_t is completely described by K Kronecker indices or row degrees, $(p_1, \ldots, p_K)$. Denote the elements of $A(L)$ and $M(L)$ as $A(L) = [\alpha_{ki}(L)]_{ki}$ and $M(L) = [m_{ki}(L)]_{ki}$. The Echelon

form imposes zero-restrictions according to

$$\begin{aligned}
\alpha_{kk}(L) &= 1 - \sum_{j=1}^{p_k} \alpha_{kk,j} L^j, \\
\alpha_{ki}(L) &= - \sum_{j=p_k - p_{ki}+1}^{p_k} \alpha_{ki,j} L^j, \quad \text{for } k \neq j, \\
m_{ki}(L) &= \sum_{j=0}^{p_k} m_{ki,j} L^j \text{ with } M_0 = A_0 \ ,
\end{aligned}$$

for $k, i = 1, \ldots, K$. The numbers p_{ki} are given by

$$p_{ki} = \begin{cases} \min\{p_k + 1, p_i\}, & \text{if } k \geq i \\ \min\{p_k, p_i\}, & \text{if } k < i \end{cases} \qquad k, i = 1, \ldots, K,$$

and denote the number of free parameters in the polynomials, $\alpha_{ki}(L),\ k \neq i$. It can be shown that this representation leads to identified parameters (see e.g. Hannan & Deistler 1988). In this setting, a measure of the overall complexity of the multiple series can be given by the McMillian degree $\sum_{j=1}^{k} p_j$ which is also the dimension of the corresponding state vector in a state space representation. Note that equal Kronecker indices, i.e. $p_1 = p_2 = \ldots = p_K$, lead to a standard, unrestricted VARMA representation.

3 Description of Estimation Methods

In the following, a description of the examined algorithms is given. Throughout, it is assumed that the data has been mean-adjusted prior to estimation. I do not distinguish between raw data and mean-adjusted data for notational ease. Most of the algorithms are discussed based on the general representation (1) and it is assumed that restrictions are imposed on the parameter vector of the VARMA model according to the Echelon form. The observed sample is $y_1, y_2, \ldots, y_T$. The vector of total parameters is denoted by β $(K^2(p+q) \times 1)$ and the vector of free parameters by γ $(n_\gamma \times 1)$. Let $\mathbf{A} := [A_1, \ldots, A_p]$ and $\mathbf{M} := [M_1, \ldots, M_q]$ be matrices

collecting the autoregressive and moving-average coefficient matrices, respectively. Define

$$\beta := \text{vec}[I_K - A_0, \mathbf{A}, \mathbf{M}],$$

where vec denotes the operator that transforms a matrix to a column vector by stacking the columns of the matrix below each other. This particular order of the free parameters allows to formulate many of the following estimation methods as standard linear regression problems. To consider zero and equality restrictions on the parameters, define a $((K^2(p+q)) \times n_\gamma)$ matrix R such that $\beta = R\gamma$. This notation is equivalent to the explicit formulation of restrictions on β such as $C\beta = c$ for suitable matrices C and c.

Hannan-Rissanen Method (HR): This is the simplest method. The procedure is easy to implement and is sometimes called the Hannan-Rissanen method or Durbin's method (Durbin 1960, Hannan & Rissanen 1982) because it corresponds to the second stage of the method proposed in Hannan & Rissanen (1982) for univariate models. See Hannan & Kavalieris (1984*b*) for the extension to the multivariate case. Recently, the estimator's asymptotic distribution in the vector case have been derived by Dufour & Jouini (2005) under quite general conditions. The idea is to use the infinite VAR representation in (4) in order to estimate the residuals u_t in a first step. In finite samples, a good approximation is a finite-order VAR, provided that the process is of low order and the roots of the moving-average polynomial are not too close to unity in modulus. The first step of the algorithm consists of a preliminary long autoregression of the type

$$y_t = \sum_{i=1}^{n_T} \Pi_i y_{t-i} + u_t, \tag{5}$$

where n_T is the lag length that is required to increase with the sample size, T. In the second stage, the residuals from (5), $\hat{u}_t^{(0)}$, $t = n_T + 1, \ldots, T$, are plugged in (1). After rearranging

(1), one gets

$$\begin{aligned} y_t &= (I_K - A_0)[y_t - \hat{u}_t^{(0)}] + A_1 y_{t-1} + \ldots + A_p y_{t-p} \\ &\quad + M_1 \hat{u}_{t-1}^{(0)} + \ldots + M_q \hat{u}_{t-q}^{(0)} + u_t, \end{aligned} \tag{6}$$

where $A_0 = M_0$ has been used. Write the above equation compactly as

$$y_t = [I_K - A_0, \mathbf{A}, \mathbf{M}] Y_{t-1}^{(0)} + u_t,$$

where $Y_{t-1}^{(0)} := [(y_t - \hat{u}_t^{(0)})', y_{t-1}', \ldots, y_{t-p}', (\hat{u}_{t-1}^{(0)})', \ldots, (\hat{u}_{t-q}^{(0)})']'$. Collecting all observations we get

$$Y = [I_K - A_0, \mathbf{A}, \mathbf{M}] X^{(0)} + U, \tag{7}$$

where $Y := [y_{n_T+m+1}, \ldots, y_T]$, $U := [u_{n_T+m+1}, \ldots, u_T]$ is the matrix of regression errors, $X^{(0)} := [Y_{n_T+m}^{(0)}, \ldots, Y_{T-1}^{(0)}]$ and $m := \max\{p, q\}$. Thus, the regression is started at $n_T + m + 1$. One could also start simply at $m + 1$, setting the initial errors to zero but we have decided not to do so. Vectorizing equation (7) yields

$$\mathrm{vec}(Y) = (X^{(0)\prime} \otimes I_K) R\gamma + \mathrm{vec}(U),$$

and the HR estimator is defined as

$$\tilde{\gamma} = [R'(X^{(0)} X^{(0)\prime} \otimes (\widehat{\Sigma}^{(0)})^{-1}) R]^{-1} R'(X^{(0)} \otimes (\widehat{\Sigma}^{(0)})^{-1}) \mathrm{vec}(Y), \tag{8}$$

because $E[\mathrm{vec}(U)\mathrm{vec}(U)'] = I_T \otimes \Sigma$ and $\widehat{\Sigma}^{(0)} = 1/T \sum \hat{u}_t^{(0)} (\hat{u}_t^{(0)})'$ is the estimated covariance matrix of the residuals. The corresponding estimated matrices are denoted by $\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_p$ and $\tilde{M}_1, \tilde{M}_2 \ldots, \tilde{M}_q$, respectively.

In their framework, Dufour & Jouini (2005) derive consistency of the parameter estimators for n_T increasing at a rate below $T^{1/2}$, i.e. $n_T \to \infty \;\; n_T/T^{1/2} \to 0$ as $T \to \infty$, and asymptotic normality for a rate below $T^{1/4}$. For univariate and multivariate models, different selection

rules for the lag length of the initial autoregression have been proposed. For example, Hannan & Kavalieris (1984*a*) propose to select n_T by AIC or BIC, while Koreisha & Pukkila (1990) propose choosing $n_T = \sqrt{T}$ or $n_T = 0.5\sqrt{T}$. In general, choosing a higher value for n_T increases the risk of obtaining non-invertible or non-stationary estimated models (Koreisha & Pukkila 1990). Throughout, we employ $n_T = 0.5\sqrt{T}$.

Three technical details are worth mentioning at this point. First, it can happen that the estimated autoregressive model (5) is not stationary. In this case, a Yule-Walker estimator is employed (see e.g. Lütkepohl 2005, 3.3). Second, the estimated VARMA model might not be invertible. We use a procedure proposed by Lippi & Reichlin (1994) in a different context in order to obtain the corresponding invertible representation. A detailed account is given in the appendix. Third, the estimated VARMA model might not be stable. In this case,different lag orders n_T in (5) are tried to obtain a stationary and invertible estimated VARMA model.

Hannan-Kavalieris-Procedure (HK): This method adds a third stage to the procedure just described. It goes originally back to Hannan & Kavalieris (1984*b*) for multivariate processes. See also Hannan & Deistler (1988, 6.5-6.7) for an extensive discussion. In contrast to HR, the resulting estimator is asymptotically efficient for Gaussian innovations. It is a Gauss-Newton procedure to maximize the likelihood function conditional on $y_t = 0,\ u_t = 0$ for $t \leq 0$ but its first iteration has sometimes been interpreted as a three-stage least squares procedure (Dufour & Pelletier (2008)). The method is computationally very easy to implement because of its recursive nature. Corresponding to the estimates of the HR algorithm, new residuals, ε_t $(K \times 1)$, are formed. One step of the Gauss-Newton iteration is performed starting from these estimates. Thus, given the output of the HR procedure, one calculates

series, ξ_t $(K \times 1)$, η_t $(K \times 1)$ and $\hat{X}_t$ $(K \times n_\gamma)$ according to

$$\begin{aligned}
\varepsilon_t &= \tilde{A_0}^{-1} \left(\tilde{A}_0 y_t - \sum_{j=1}^{p} \tilde{A}_j y_{t-j} - \sum_{j=1}^{q} \tilde{M}_j \varepsilon_{t-j} \right), \\
\xi_t &= \tilde{A_0}^{-1} \left(- \sum_{j=1}^{q} \tilde{M}_j \xi_{t-j} + \varepsilon_t \right), \\
\eta_t &= \tilde{A_0}^{-1} \left(- \sum_{j=1}^{q} \tilde{M}_j \eta_{t-j} + y_t \right), \\
\hat{X}_t &= \tilde{A_0}^{-1} \left(- \sum_{j=1}^{q} \tilde{M}_j \hat{X}_{t-j} + (\tilde{Y}_t' \otimes I_K) R \right),
\end{aligned}$$

for $t = 1, 2, \ldots, T$ and $y_t = \varepsilon_t = \xi_t = \eta_t = 0_{K \times 1}$ and $\hat{X}_t = 0_{K \times n_\gamma}$ for $t \leq 0$ and $\tilde{Y}_t$ is structured as $Y_t^{(0)}$ with ε_t in place of $\hat{u}_t^{(0)}$. Given these quantities, we compute the HK estimate as

$$\hat{\gamma} = \left(\sum_{m+1}^{T} \hat{X}_{t-1}' \widehat{\Sigma}_t^{-1} \hat{X}_{t-1} \right)^{-1} \left(\sum_{m+1}^{T} \hat{X}_{t-1} \widehat{\Sigma}^{-1} (\varepsilon_t + \eta_t - \xi_t) \right),$$

where $\widehat{\Sigma} := T^{-1} \sum \varepsilon_t \varepsilon_t'$, $m := \max\{p, q\}$ as before and the estimated coefficient matrices are denoted by $\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_p$ and $\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_q$, respectively.

Hannan & Kavalieris (1984*b*) showed consistency and asymptotic normality of these estimators. It is possible to use this procedure iteratively, starting the above recursions in the second iteration with the newly obtained parameter estimates from the HK procedure, and so on until convergence.

Generalized Least Squares (KP): Also this procedure has three stages. Koreisha & Pukkila (1990) proposed the method for univariate ARMA models and Kavalieris, Hannan & Salau (2003) proved efficiency of the KP estimates in this case. The motivation is the same as for the HR estimator. Given consistent estimates of the residuals, we can estimate the parameters of the VARMA representation by least squares. However, Koreisha & Pukkila (1990) note that in finite samples the residuals are estimated with error. This implies that the actual regression error is serially correlated in a particular way due to the structure of

the underlying VARMA process. The KP procedure tries to take this into account. Similar approaches have been proposed by Flores de Frutos & Serrano (2002) and Choudhury & Power (1998)

I consider a multivariate generalization of the three-stage procedure of Koreisha & Pukkila (1990). In the first stage, preliminary estimates of the innovations are obtained by a long autoregression as in (5). Koreisha & Pukkila (1990) *assume* that the residuals obtained from (5) correspond to the true residuals up to an uncorrelated error term, $u_t = \hat{u}_t^{(0)} + \epsilon_t$. If this expression is inserted in (1), one obtains

$$\begin{aligned} A_0 y_t &= \sum_{j=1}^{p} A_j y_{t-j} + A_0(\hat{u}_t^{(0)} + \epsilon_t) + \sum_{j=1}^{q} M_j(\hat{u}_{t-j}^{(0)} + \epsilon_{t-j}), \\ y_t - \hat{u}_t^{(0)} &= (I - A_0)(y_t - \hat{u}_t^{(0)}) + \sum_{j=1}^{p} A_j y_{t-j} \\ &\quad + \sum_{j=1}^{q} M_j \hat{u}_{t-j}^{(0)} + \zeta_t. \qquad (9) \\ \zeta_t &= A_0 \epsilon_t + \sum_{j=1}^{q} M_j \epsilon_{t-j} \qquad (10) \end{aligned}$$

Thus, the error term, ζ_t, in a regression of $y_t - \hat{u}_t^{(0)}$ on its lagged values and the estimated residuals is a moving-average process of order q. Thus, a least squares regression in (9) is not efficient. In the second stage, one estimates the coefficients in (9) by ordinary least squares: Let $z_t := y_t - \hat{u}_t^{(0)}$ and $Z := [z_{n_T+m+1}, \ldots, z_T]$. The second stage estimate is given analogously to the HR final estimate by

$$\tilde{\tilde{\gamma}} = [R'(X^{(0)}X^{(0)\prime} \otimes I_K)R]^{-1} R'(X^{(0)} \otimes I_K)\mathrm{vec}(Z),$$

and the residuals are computed in the usual way, that is

$$\tilde{\tilde{\zeta}}_t = z_t - (Y_{t-1}^{(0)\prime} \otimes I_K)R\tilde{\tilde{\gamma}}.$$

The covariance matrix of these residuals, Σ_ζ, is estimated as usual. From (10) one obtains

the covariance matrix of the approximation error as

$$\text{vec}(\tilde{\Sigma}_\epsilon) = \left(\sum_{i=0}^{q} (\tilde{\tilde{M}}_i \otimes \tilde{\tilde{M}}_i) \right)^{-1} \text{vec}(\tilde{\Sigma}_\zeta),$$

where the $\tilde{\tilde{M}}_j$ are formed from the corresponding elements in $\tilde{\tilde{\gamma}}$. These estimates are then used to build the covariance matrix of $\zeta = (\zeta'_{n_T+m+1} \dots \zeta'_T)'$. Let $\Phi := E[\zeta\zeta']$ and denote its estimate by $\hat{\Phi}$. In the third stage, we re-estimate (9) by GLS as

$$\hat{\tilde{\gamma}} = [R'(X^{(0)} \otimes I_K)\hat{\Phi}^{-1}(X^{(0)'} \otimes I_K)R]^{-1} R'(X^{(0)} \otimes I_K)\hat{\Phi}^{-1}\text{vec}(Z).$$

In comparison to the HR estimator, the main difference lies in the weighting with $\hat{\Phi}^{-1}$.

Iterative Least Squares (IHR) Kapetanios (2003) suggested to use the HR algorithm iteratively. The parameter estimates of the HR algorithm are employed to construct new residuals which can be used to perform another least squares operation. Denote the estimate of the HR procedure by $\tilde{\gamma}^{(1)}$. We may obtain new residuals by

$$\text{vec}(\hat{U}^{(1)}) = \text{vec}(Y) - (X^{(0)'} \otimes I_K)R\tilde{\gamma}^{(1)}.$$

Therefore, it is possible to set up a new matrix of regressors $X^{(1)}$ that is of the same structure as $X^{(0)}$ but uses the newly obtained estimates of the residuals $\hat{u}_t^{(1)}$ in $\hat{U}^{(1)}$. Generalized least squares as in (8) in

$$\text{vec}(Y) = (X^{(1)'} \otimes I_K)R\gamma + \text{vec}(U)$$

yields a new estimate $\tilde{\gamma}^{(2)}$. Denote the vector of estimated residuals at the i^{th} iteration by $\hat{U}^{(i)}$. Then we iterate least squares regressions until $||\text{vec}(\hat{U}^{(i)}) - \text{vec}(\hat{U}^{(i-1)})||$ becomes small relative to $||\text{vec}(\hat{U}^{(i-1)})||$, where $||.||$ is some norm. According to Kapetanios (2003), the IHR algorithm is consistent and has the same asymptotic properties as the HR method. In contrast to the above-mentioned regression-based procedures, the IHR procedure is iterative

but the computational load is still small.

Maximum Likelihood Estimation (MLE): The dominant approach to the estimation of VARMA models has been of course maximum likelihood estimation. The exact likelihood of a VARMA (p, q) model was first derived by Hillmer & Tiao (1979) and Nicholls & Hall (1979).[2] The presentation here is summarizing the derivation of the exact likelihood as described in Reinsel (1993, 5.3). Given the sample, $y_1, ..., y_T$, and assuming that the innovations u_t are normally distributed, one can summarize equation (1) by defining $y := (y_1', \ldots, y_T')'$, $u := (u_1', \ldots, u_T')'$ and $y^0 := (y_{-p+1}', \ldots, y_0', u_{-q+1}', \ldots, u_0')'$ and writing

$$\mathfrak{A}y \quad = \quad \mathfrak{A}_0 y_0 + \mathfrak{M}u,$$

where $\mathfrak{A}$, $\mathfrak{A}_0$, $\mathfrak{M}$ are functions of $A_0, A_1, \ldots, A_p$ and $M_1, \ldots, M_q$, see Reinsel (1993). Since the u_t are Gaussian and $y_t = A(L)^{-1}M(L)u_t$, all terms in y_0 as well as in u are Gaussian too and y_0 and u are independent. Thus, y, conditional on the parameters, is Gaussian as well with

$$y \sim N(0, \mathfrak{A}^{-1}(\mathfrak{A}_0 E[y_0 y_0']\mathfrak{A}_0' + \mathfrak{M}(I_T \otimes \Sigma)\mathfrak{M}')\mathfrak{A}'^{-1}).$$

Denote the covariance by $\Gamma_0 := \mathfrak{A}^{-1}(\mathfrak{A}_0 E[y_0 y_0']\mathfrak{A}_0' + \mathfrak{M}(I_T \otimes \Sigma)\mathfrak{M}')\mathfrak{A}'^{-1}$. The log likelihood function of the vector of free parameters, γ, and the covariance matrix of the residuals can be expressed as

$$\ell(\gamma, \Sigma) \propto -1/2 \ln |\Gamma_0| - \frac{1}{2} y' (\Gamma_0)^{-1} y, \tag{11}$$

where the dependence of Γ_0 on γ and Σ is omitted on the right hand side. The formulation makes clear that the maximization of (11) is highly nonlinear. Exact maximum likelihood estimation "backcasts" the initial values in that the term $E[y_0 y_0']$ needs to be calculated. The procedure is implemented using the formulation of the exact likelihood by Mauricio (1995) as implemented in GAUSS 9.0 and the sqpSolveMT function is used to maximize it. The starting values are the true parameter values and a limited number of iterations is allowed

[2] See also Deistler & Pötscher (1984) on the behavior of the likelihood function for ARMA models.

and therefore the results from the exact maximum likelihood procedure must be regarded as a benchmark rather than as a realistic estimation alternative.

Vector Autoregressive Approximations (VAR) An alternative to VARMA modeling is using just a pure autoregressive model - as it is very often done in practice. As there is no true lag order, we employ the AIC and BIC information criteria to choose a lag length. The corresponding VARs are denoted by VAR(AIC) and VAR(BIC). This is done in order to assess the potential merits of VARMA modeling compared to standard VAR modeling.

4 Monte Carlo Study

I compare the performance of different estimation methods using a variety of measures. The parameter estimation precision, the accuracy of point forecasts and the precision of the estimated impulse responses are compared. These measures are related. For instance, one would expect that an algorithm that yields accurate parameter estimates performs also well in a forecasting exercise. However, almost all results on the efficiency of different estimators rely on asymptotic theory. There is no guarantee that the ranking of estimators based on large samples is the same in small samples. This phenomenon is simply due to the limited information in small samples. While it is not clear a priori whether there are important differences with respect to the different measures used, it is worth investigating these issues separately in order to uncover potential advantages or disadvantages of the algorithms.

Apart from the performance measures mentioned above, I am also interested in the "technical reliability" of the algorithms. This is not a trivial issue as the results will make clear. First, I consider the number of cases when the algorithms yielded non-stationary models. Second, the number of cases when the models yielded extreme outliers is counted. For the IHR algorithm another relevant statistic is the number of replications where the iterations did not converge. For all algorithms, the estimates of the HR procedure are adopted in the case that a particular algorithm fails for one of the mentioned reasons.

I consider various processes and variations of them as described below. For all data generating processes, I simulate $N = 1000$ series of length $T = 100$. The sample size could be

regarded as typical for macroeconomic time series applications.[3] I consider mostly processes that have been used in the literature to demonstrate the virtue of specific algorithms but I also consider examples taken from estimated processes.

4.1 Performance Measures

4.1.1 Parameter Estimates

The accuracy of the different parameter estimators are compared. The parameters may be of independent interest to the researcher. Denote by $\hat{\gamma}_{\mathcal{A},n}$ the estimate of γ obtained by some algorithm $\mathcal{A}$ at the nth replication of the simulation experiment. The accuracy of an estimator is summarized as the trace of the estimated MSE matrix

$$\text{tr } MSE_{\mathcal{A}} = \text{tr } \left(\frac{1}{N} \sum_{n=1}^{N} (\hat{\gamma}_{\mathcal{A},n} - \gamma)(\hat{\gamma}_{\mathcal{A},n} - \gamma)' \right).$$

The index n refers to a particular replication of the simulation experiment, $n = 1, \ldots, N$. In the graphs, we compute the ratio of the trace of the MSE matrix of a particular algorithm relative to those of the MLE method

$$\text{tr } MSE_{\mathcal{A}} / \text{tr } MSE_{MLE}.$$

4.1.2 Forecasting

Forecasting is one of the main objectives in time series modeling. To assess the forecasting power of different VARMA estimation algorithms, the traces of forecast mean squared error (FMSE) matrices of 1-step and 4-step-ahead out-of-sample forecasts are compared. Specifically, the trace of the FMSE matrix at horizon h for the algorithm $\mathcal{A}$ is

$$\text{tr } FMSE_{\mathcal{A}}(h) = \text{tr } \left(\frac{1}{N} \sum_{n=1}^{N} (y_{T+h,n} - \hat{y}_{T+h|T,n})(y_{T+h,n} - \hat{y}_{T+h|T,n})' \right),$$

[3] In an earlier version of the paper, the algorithms were also compared for a sample size of $T = 200$. The results, however, did not differ much and are therefore omitted.

where $y_{T+h,n}$ is the value of y_t at $T+h$ for the nth replication and $\hat{y}_{T+h|T,n}$ denotes the corresponding h-step ahead forecast at origin T and the dependence on $\mathcal{A}$ is suppressed on the right hand side. For given estimated parameters and a finite sample at hand, the corresponding estimate of the white noise sequence, $\hat{u}_t$, is used to compute forecasts recursively according to

$$\hat{y}_{T+h|T} = A_0^{-1}\left(\sum_{j=1}^{p} A_j \hat{y}_{T+h-j|T} + \sum_{j=h}^{q} M_j \hat{u}_{T+h-j}\right),$$

for $h = 1, \ldots, q$. For $h > q$, the forecast is simply $\hat{y}_{T+h|T} = A_0^{-1}\sum_{j=1}^{p} A_j \hat{y}_{T+h-j|T}$. The forecast precision of an algorithm $\mathcal{A}$ is measured relative to the MLE method

$$\text{tr } FMSE_{\mathcal{A}}(h)/\text{tr } FMSE_{\text{MLE}}(h).$$

4.1.3 Impulse Response Analysis

Researchers might also be interested in the accuracy of the estimated impulse response function as in (3),

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} = \Phi(L)u_t,$$

since it displays the propagation of shocks to y_t over time. I compute impulse response mean squared errors (IRMSE) at two different horizons, $h = 1$ and $h = 4$. Let $\psi_h = \text{vec}(\Phi_h)$ denote the vector of responses of the system to shocks h periods ago. A measure of the accuracy of the estimated impulse responses is

$$\text{tr } IRMSE_{\mathcal{A}}(h) = \text{tr}\left(\frac{1}{N}\sum_{n=1}^{N}(\psi_h - \hat{\psi}_{h,n})(\psi_h - \hat{\psi}_{h,n})'\right),$$

where $\hat{\psi}_{h,n}$ is the estimated response. The precision of the estimated responses for a particular algorithm are again measured relative to the precision of the MLE method

$$\frac{\text{tr } IRMSE_{\mathcal{A}}(h)}{\text{tr } IRMSE_{\text{MLE}}(h)}.$$

4.2 Generated Systems

4.2.1 Small-Dimensional Systems

DGP I: The first two-dimensional process is a simplified version of the process fitted by Reinsel (1993, pp. 253-255) to U.S. business investment and inventories data. It is a bivariate VARMA(2,2) model with Kronecker indices $(p_1, p_2) = (2, 1)$. Precisely, the process is given by

$$\begin{aligned}\begin{pmatrix} 1 & 0 \\ 0.4 & 1 \end{pmatrix} y_t &= \begin{pmatrix} 0.51 & 0 \\ 0.52 & \alpha_{22,1} \end{pmatrix} y_{t-1} + \begin{pmatrix} -0.13 & 0 \\ 0 & 0 \end{pmatrix} y_{t-2} \\ &\quad + \begin{pmatrix} 1 & 0 \\ 0.4 & 1 \end{pmatrix} u_t + \begin{pmatrix} 0 & m_{21,1} \\ 0 & 0 \end{pmatrix} u_{t-1} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} u_{t-2},\end{aligned}$$

and

$$\Sigma = \begin{pmatrix} 4.97 & \\ 1.69 & 15.96 \end{pmatrix}.$$

This admittedly small process is used here to check the performance of the algorithms in the case $A_0 \neq I_K$. In addition, many parameter values are zero even though the process is identified.

The autoregressive polynomial has three non-zeros eigenvalues and the moving-average polynomial has one eigenvalue different from zero. Denote the eigenvalues of the autoregressive and moving-average part by λ^{ar} and λ^{ma}, respectively. These eigenvalues are varied and the parameters $\alpha_{12,1}$, $m_{21,1}$ are set accordingly. For this and the following DGPs, I consider parameterizations with medium eigenvalues (MEV), large positive autoregressive

eigenvalues ($LPAREV$), large negative autoregressive eigenvalues ($LNAREV$), large positive moving-average eigenvalues ($LPMAEV$) and large negative moving-average eigenvalues ($LNMAEV$). The parameter values corresponding to the different parameterizations can be found in Table 1 for all DGPs.

For the present process the MEV parametrization corresponds to $\alpha_{22,1} = 0.66$, $m_{21,1} = -0.13$ with eigenvalues $\lambda_1^{ar} = 0.255-0.25i,\ \lambda_2^{ar} = 0.255+0.25i,\ \lambda_3^{ar} = 0.66$ and $\lambda_1^{ma} = -0.052$. I fit restricted VARMA models in Echelon form to the simulated data.

DGP II: The second DGP is based on an empirical example taken from Lütkepohl (2005). A VARMA(2,2) model is fitted to West-German income and consumption data. The variables were the first differences of log income, y_1, and log consumption, y_2. More specifically, a VARMA (2,2) model with Kronecker indices $(p_1, p_2) = (0, 2)$ was assumed such that

$$\begin{aligned} y_t &= \begin{pmatrix} 0 & 0 \\ 0 & \alpha_{22,1} \end{pmatrix} y_{t-1} + \begin{pmatrix} 0 & 0 \\ 0 & \alpha_{22,2} \end{pmatrix} y_{t-2} + u_t \\ &\quad + \begin{pmatrix} 0 & 0 \\ 0.31 & m_{22,1} \end{pmatrix} u_{t-1} + \begin{pmatrix} 0 & 0 \\ 0.14 & m_{22,2} \end{pmatrix} u_{t-2} \end{aligned}$$

and

$$\Sigma = \begin{pmatrix} 1.44 & \\ 0.57 & 0.82 \end{pmatrix} \times 10^{-4}.$$

While the autoregressive part has two distinct, real roots, the moving-average polynomial has two complex-conjugate roots in the original specification. We vary again some of the parameters in order to obtain different eigenvalues. In particular, we maintain the property that the process has two complex moving-average eigenvalues which are less than one in modulus.

The MEV parametrization corresponds to the estimated process with $\alpha_{22,1} = 0.23$, $\alpha_{22,2} = 0.06$, $m_{22,1} = -0.75$ and $\hat{m}_{22,2} = 0.16$. These values imply the following eigenvalues $\lambda_1^{ar} = 0.385\ \lambda_2^{ar} = -0.159,\ \lambda_1^{ma} = -0.375 + 0.139i,\ \lambda_2^{ma} = -0.375 - 0.139i$. VARMA

models with restrictions given by the Kronecker indices were used.

4.2.2 Higher-Dimensional Systems

DGP III: I consider a three-dimensional system that was used extensively in the literature by e.g. Koreisha & Pukkila (1989), Flores de Frutos & Serrano (2002) and others for illustrative purposes. Koreisha & Pukkila (1989) argue that the chosen model is typical for real data applications in that "[...]the density of nonzero elements is low, the variation in magnitude of parameter values is broad and the feedback/causal mechanisms are complex.". The data is generated according to

$$y_t = \begin{pmatrix} \alpha_{11,1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0.4 & 0 \end{pmatrix} y_{t-1} + u_t + \begin{pmatrix} 0 & 1.1 & 0 \\ 0 & -0.6 & 0 \\ 0 & 0 & m_{33,1} \end{pmatrix} u_{t-1}$$

and

$$\Sigma = \begin{pmatrix} 1 & & \\ -0.7 & 1 & \\ 0.4 & 0 & 1 \end{pmatrix}.$$

The Kronecker indices are given by $(p_1, p_2, p_3) = (1, 1, 1)$ and corresponding VARMA$(1, 1)$ models are fitted to the data. While this DGP is of higher dimension, the parameter matrices are more sparse. This property is reflected in the fact that the autoregressive polynomial and the moving-average polynomial have both few eigenvalues different from zero.

The parameters $\alpha_{11,1}$ and $m_{33,1}$ are varied in order to generate particular eigenvalues of the autoregressive and moving-average polynomials as in the foregoing examples. The MEV specification corresponds to the process used in Koreisha & Pukkila (1989) and has eigenvalues $\lambda^{ar} = 0.7$ and $\lambda_1^{ma} = -0.6$ and $\lambda_2^{ma} = 0.5$.

DGP IV: This process has been used in the simulation studies of Koreisha & Pukkila (1987). The process is similar to the DGP III. Here it is used in particular to investigate the

performance of the algorithms for the case of high-dimensional systems. The five variables are generated according to the following VARMA (1,1) structure

$$y_t = \begin{pmatrix} \alpha_{11,1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0 & -0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 0 \end{pmatrix} y_{t-1} + u_t + \begin{pmatrix} 0 & 0 & 0 & -1.1 & 0 \\ 0 & 0 & 0 & 0 & -0.2 \\ 0 & 0 & 0 & 0 & 0 \\ 0.55 & 0 & 0 & -0.8 & 0 \\ 0 & 0 & 0 & 0 & m_{55,1} \end{pmatrix} u_{t-1}$$

and

$$\Sigma = \begin{pmatrix} 1 & & & & \\ 0.2 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0.7 & 1 & \\ 0 & 0 & 0 & -0.4 & 1 \end{pmatrix}.$$

The true Kronecker indices are $(p_1, p_2, p_3, p_4, p_5) = (1, 1, 1, 1, 1)$ and corresponding VARMA models in Echelon form are fitted to the data. The MEV parametrization corresponds to the one used by Koreisha & Pukkila (1987). That is, $\alpha_{11,1} = 0.5$ and $m_{55,1} = -0.6$ with eigenvalues $\lambda_1^{ar} = 0.5$, $\lambda_2^{ar} = 0 + i0.57$, $\lambda_3^{ar} = 0 - i0.57$, $\lambda_1^{ma} = -0.6$, $\lambda_2^{ma} = -0.4 + i0.67$ and $\lambda_3^{ma} = -0.4 - i0.67$.

4.3 Results

The results are summarized in Table 2 and Figures 1 to 4. The table shows the frequency of cases when the algorithms failed for different reasons. The figures plot the various MSE ratios discussed above.

Table 2 displays the frequency of cases in which the algorithms yielded models that were not stationary, yielded extreme outliers or, in the case of the IHR algorithm, did not converge. The IHR algorithm is regarded as non-convergent if it did not converge after 500 iterations. As expected, the algorithms yield non-stationary models more frequently when the eigenvalues of

the autoregressive polynomial are close to one in absolute value but also when the eigenvalues of the moving-average part become close to the non-invertibility region. The algorithms yield non-stationary models most often for DGP I and much less frequently for the other DGPs. The HK algorithm fails most often in this respect and especially for DGP I even when the eigenvalues of the autoregressive part are well inside the stationary region. The most reliable algorithms are HR and KP. The algorithms almost never yielded parameter estimates which were extremely different from their true values. The convergence properties of the IHR algorithm depend stronger on the simulated DGP than on the chosen parametrization. The algorithm does not converge very often for the DGPs I and II which are more restricted by the Echelon form. The problem is aggravated by large eigenvalues of the moving-average part. In sum, the reliability of certain algorithms depends primarily on the structure of the simulated DGP. The parameterizations are of minor importance and HR and KP are very reliable irrespective of the DGP and parameterization.

With respect to parameter estimation accuracy, the differences between the algorithms are generally more pronounced when the moving-average polynomial has eigenvalues that are close to one in absolute value. The HR algorithm delivers the most precise forecasts for DGP I for small moving-average values but is almost dominated by the other algorithms for DGPs II and III while its performance is average for DGP IV. The relative performance of the KP estimator varies considerably between DGPs. The algorithm's parameter estimates are quite precise for DGP I LPMAEV and LNMAEV and for DGP II. The KP estimator is, however, the worst for DGP IV. Thus, it is competitive only for the small dimensional processes in our study. The IHR estimator delivers the most precise parameter estimates for the MEV parameterizations for DGP II and IV but otherwise yields parameter estimates whose precision are close to the average of the investigated methods. Thus, in this respect, the IHR estimator is relatively robust across DGPs. The HK estimator is worse in terms of parameter estimation for DGP I but otherwise does quite well in relative terms. In particular, the HK method is much better when the number of estimated parameters increases, that is for DGP III and IV. Summarizing, the HK procedure is overall the best alternative to MLE despite the high number of cases when the algorithm yielded non-stationary models or outliers

in the case of DGP I. Nevertheless, even the best alternative can be quite imprecise compared to MLE. This does not necessarily mean that HK is not a relatively good estimator because the MLE procedure starts with the true parameter values and therefore the procedure represents an ideal case in this context.

The differences in terms of forecasting precision are less pronounced. Additionally, even though some algorithms estimate the parameters more accurately than others, they are not necessarily superior in terms of forecasting accuracy. The ranking might change. Apart from DGP I, the VARMA algorithms do better than the VARs specified by AIC or BIC. However, given that the orders of the VARMA models are fixed and correspond to the true orders, the comparison is biased in favor of VARMA modeling. Increasing the forecast horizon does generally reduce the differences between the algorithms. An exception is the LNMAEV case in which the HR estimator is performing poorly at $h = 4$. Increasing the complexity in terms of Kronecker indices does have minor effects. The HR estimator yields usually comparable but sometimes slightly worse forecasts than the other VARMA algorithms. However, it performs often poorly in the case of large eigenvalues in the moving-average part. The KP and the IHR procedure do quite well in forecasting depending on the specific DGP and number of observations. The HK procedure, however, seems to be slightly preferable. Apart from DGP I, the HK procedure is often superior to the other algorithms and, in any case, close to the best-performing method. The MLE benchmark is always superior to all simple algorithms apart from one case, DGP III, LNMAEV. In general, however, the differences are small, in particular in comparison to the rather large differences in terms of parameter estimation accuracy. For the simulated processes, HK is a good alternative algorithm to MLE if forecasting is the objective.

The precision of the estimated impulse responses varies much more between the algorithms. In most cases the VARMA algorithms do comparably or better than the VAR approximations but, as mentioned above, this comparison is biased in favor of the use of VARMAs. When the impulse response horizon is increased, VARMA modeling becomes much more advantageous in comparison with the VAR approximations. At short horizons the picture is rather mixed depending on the algorithms and DGPs. For the rather simple

DGP I, there are few advantages from VARMA modeling. For the other DGPs there are in principle considerable advantages in particular for $h = 4$. The HR algorithm estimates the impulse responses with comparable or slightly worse accuracy than the other VARMA algorithms. The precision of the impulse response estimates obtained by KP are typically on average. The IHR algorithm is performing comparably or slightly better than the HR and KP algorithms. Apart from DGP I, HK is often the preferable method, in particular for short horizons. Generally, the impulse response estimates obtained by MLE are much more precise than the corresponding estimates obtained by the other algorithms. These results correspond to the statements made above about the algorithms' relation in terms of parameter estimation accuracy. Overall, VARMA modeling turns out to be potentially quite advantageous if one is interested in the impulse responses of the DGP. The precision obtained by MLE is, however, rarely obtained by any of the simpler VARMA estimation algorithms.

5 Conclusion

Despite the theoretical advantages of VARMA models compared to simpler VAR models, they are rarely used in applied macroeconomic work. While Gaussian maximum likelihood estimation is theoretically attractive, it is plagued with various numerical problems. Therefore, simpler estimation algorithms are compared in this paper by means of a Monte Carlo study. The evaluation criteria used are the precision of the parameter estimates, the accuracy of point forecasts and the accuracy of the estimated impulse responses. The VARMA algorithms are also compared to two benchmark VARs in order to judge the potential merits of VARMA modeling.

It has been shown in the simulations that VARMA modeling can be advantageous compared to VAR modeling. While the advantages are potentially minor with respect to forecasting precision, the results suggest that the impulse responses can be estimated more accurately by using VARMA models, provided that the model is specified correctly. There can be large differences between the algorithms. Overall, the algorithm of Hannan & Kavalieris (1984*b*) which is closest to maximum likelihood estimation seems to be superior to the other simple

estimation algorithms in terms of all three criteria. In particular, when the complexity of the simulated systems increases. A concern, however is the instability and poor performance of the algorithm for some DGPs. Thus, one might prefer to combine the results from different estimation algorithms.

While this study suggests that there are potentially considerable gains from VARMA modeling, a reliable, accurate as well as computationally efficient algorithm for the estimation of VARMA models still remains to be developed. The results imply that this algorithm should be close to a robust maximum likelihood method. Such an algorithm would have to be able to deal with various issues which are not considered in this study. The algorithm should work well in the case of integrated and cointegrated multivariate series. The algorithm must give reasonable results with extremely over-specified processes as well as in the presence of various data irregularities such as outliers, structural breaks etc. The applicability of such an algorithm would also crucially depend on the existence of a reliable specification procedure. These topics, however, are left for future research.

References

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Trans. Autom. Control AC-19* pp. 716–723.

Bauer, D. (2005), 'Estimating linear dynamical systems using subspace methods', *Econometric Theory* **21**, 181–211.

Chari, V., Kehoe, P. J. & McGrattan, E. R. (2008), 'Are structural VARs with long-run restrictions useful in developing business cycle theory?', *Journal of Monetary Economics* **55**(8), 1337–1352.

Choudhury, A. H. & Power, S. (1998), 'A simplified GLS estimator for autoregressive moving-average models', *Applied Economics Letters* **5**, 247–250.

Christiano, L. J., Eichenbaum, M. & Vigfusson, R. (2006), Assessing structural VARs, *in* 'NBER Macroeconomics Annual', Vol. 21, MIT Press.

Cooley, T. F. & Dwyer, M. (1998), 'Business cycle analysis without much theory. A look at structural VARs', *Journal of Econometrics* **83**, 57–88.

Deistler, M. & Pötscher, B. M. (1984), 'The behaviour of the likelihood function for ARMA models', *Advances in Applied Probability* **16**, 843–865.

Dufour, J. M. & Jouini, T. (2005), Asymptotic distribution of a simple linear estimator for varma models in echelon form, *in* P. Duchesne & B. Rémillard, eds, 'Statistical Modeling and Analysis for Complex Data Problems', Kluwer/Springer-Verlag, New York, chapter 11, pp. 209–240.

Dufour, J. M. & Pelletier, D. (2008), 'Practical methods for modelling weak VARMA processes: Identification, estimation and specification with a macroeconomic application', *Discussion Paper, McGill University, CIREQ and CIRANO* .

Durbin, J. (1960), 'The fitting of time-series models', *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* **28**(3), 233–244.

Fernández-Villaverde, J., Rubio-Ramírez, J. F., Sargent, T. J. & Watson, M. W. (2007), 'A,B,C's (and D)'s of understanding VARs', *American Economic Review* **97**(3), 1021–1026.

Flores de Frutos, R. & Serrano, G. R. (2002), 'A Generalized Least Squares Estimation Method For VARMA Models', *Statistics* **13**(4), 303–316.

Hannan, E. J. & Deistler, M. (1988), *The Statistical Theory of Linear Systems*, Wiley, New York.

Hannan, E. J. & Kavalieris, L. (1984*a*), 'A method for autoregressive-moving average estimation', *Biometrika* **71**(2), 273–280.

Hannan, E. J. & Kavalieris, L. (1984*b*), 'Multivariate linear time series models', *Advances in Applied Probability* **16**(3), 492–561.

Hannan, E. J. & Rissanen, J. (1982), 'Recursive estimation of mixed autoregressive-moving average order', *Biometrika* **69**(1), 81–94.

Hillmer, S. C. & Tiao, G. C. (1979), 'Likelihood function of stationary multiple autoregressive moving average models', *Journal of the American Statistical Association* **74**, 652–660.

Kapetanios, G. (2003), 'A note on the iterative least-squares estimation method for ARMA and VARMA models', *Economics Letters* **79**(3), 305–312.

Kavalieris, L., Hannan, E. J. & Salau, M. (2003), 'Generalized Least Squares Estimation of ARMA Models', *Journal of Time Series Anaysis* **24**(2), 165–172.

Koreisha, S. & Pukkila, T. (1987), 'Identification of Nonzero Elements in the Polynomial Matrices of Mixed VARMA Processes', *Journal of the Royal Statistical Society. Series B* **49**(1), 112–126.

Koreisha, S. & Pukkila, T. (1989), 'Fast Linear Estimation Methods for Vector ARMA Models', *Journal of Time Series Anaysis* **10**(4), 325–339.

Koreisha, S. & Pukkila, T. (1990), 'A generalized least squares approach for estimation of autoregressive moving average models', *Journal of Time Series Analysis* **11**(2), 139–151.

Larimore, W. E. (1983), System identification, reduced order filters and modelling via canonical variate analysis, *in* H. S. Rao & P. Dorato, eds, 'Proceedings of the 1983 American Control Conference', New York: Piscataway, pp. 445–451.

Lippi, M. & Reichlin, L. (1994), 'VAR analysis, nonfundamental representations, blaschke matrices', *Journal of Econometrics* **63**(1), 307–325.

Lütkepohl, H. (1984*a*), 'Linear aggregation of vector autoregressive moving average processes', *Economics Letters* **14**(4), 345–350.

Lütkepohl, H. (1984*b*), 'Linear transformations of vector ARMA processes', *Journal of Econometrics* **26**(3), 283–293.

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

Mauricio, J. A. (1995), 'Exact maximum likelihood estimation of stationary vector ARMA models', *Journal of the American Statistical Association* **90**(429), 282–291.

Metaxoglou, K. & Smith, A. (2007), 'Maximum likelihood estimation of varma models using a state-space em algorithm', *Journal of Time Series Analysis* **28**(5), 666–685.

Newbold, P. & Granger, C. W. J. (1974), 'Experiences with forecasting univariate time series and combination of forecasts', *Journal of the Royal Statistical Society A* **137**, 131–146.

Nicholls, D. F. & Hall, A. D. (1979), 'The exact likelihood function of multivariate autoregressive-moving average models', *Biometrika* **66**(2), 259–264.

Poskitt, D. S. (1992), 'Identification of echelon canonical forms for vector linear processes using least squares', *Annals of Statistics* **20**, 196–215.

Reinsel, G. C. (1993), *Elements of Multivariate Time Series Analysis*, Springer-Verlag, New York.

Stock, J. H. & Watson, M. W. (2007), 'Why has u.s. inflation become harder to forecast?', *Journal of Money, Credit and Banking* **39**(s1), 3–33.

Van Overschee, P. & DeMoor, B. (1994), 'N4sid: Subspace algorithms for the identification of combined deterministic-stochastic processes', *Automatica* **30**(1), 75–93.

A Inversion of Moving-Average Roots

This section summarizes how to obtain an invertible moving-average representation from a non-invertible one. I draw on the results of Lippi & Reichlin (1994). Consider the general VARMA representation given in the paper

$$A(L)y_t = M(L)u_t = M(L)Pv_t = N(L)v_t, \tag{12}$$

where the matrix P is the Cholesky decomposition of the covariance matrix of u_t, i.e. $PP' = \Sigma$. We define $N(L) := M(L)P$ such that v_t is orthonormal white noise, i.e. $E[v_t v_t'] = I_K$.

Assume the moving-average polynomial is not invertible. As the roots of $M(z)$ are the same as those of $N(z)$, assume that in $|N(z)| = \tau \prod_{i=1}^{\bar{q}} (z - \lambda_i)^{m_i}$ the first root $\lambda_1 \in \mathbb{C}$ of multiplicity m_1 has modulus smaller than one, $|\lambda_1| < 1$. Therefore, $N(\lambda_1)$ is singular and there exists a non-trivial solution to the system

$$N(\lambda_1)y = 0. \tag{13}$$

Denote one such vector by g, normalized such that $g'g = 1$. Form an orthogonal matrix $\mathbf{K}$ ($\mathbf{K}\mathbf{K}' = I_K$) that contains in the first column g. This can be achieved e.g. by using a Gram-Schmidt procedure.

Consider the first column of the matrix polynomial

$$\tilde{N}(L) = N(L) \cdot \mathbf{K},$$

which consists of the polynomials $\tilde{n}_{k1}(z)$, $k = 1, \ldots, K$. Then $(\tilde{n}_{11}(\lambda_1), \ldots, \tilde{n}_{K1}(\lambda_1))' = N(\lambda_1)g = 0_K$, i.e. λ_1 is a root in all polynomials in the first column. Therefore, multiplication by the Blaschke matrix

$$R_{\lambda_1}^{-1}(L) := \begin{pmatrix} \frac{1-\bar{\lambda}_1 L}{L-\lambda_1} & 0 \\ 0 & I_{K-1} \end{pmatrix},$$

where $\bar{\lambda}_1$ denotes the conjugate of λ_1, yields again a finite-order polynomial

$$\hat{N}(L) = N(L)\mathbf{K}R_{\lambda_1}^{-1}(L) = \hat{N}_0 + \hat{N}_1 L + \ldots + \hat{N}_q L^q$$

and λ_1 is a root of $\hat{N}(z)$ of multiplicity $m_1 - 1$. That is, if $m_1 = 1$ and λ_1 is the only root with modulus smaller than one, then $\hat{N}(L)$ is invertible.

Thus, we can transform the original representation (12) as

$$\begin{aligned} A(L)y_t &= N(L)v_t, && (14)\\ &= N(L)\mathbf{K}R_{\lambda_1}^{-1}(L)R_{\lambda_1}(L)\mathbf{K}'v_t, && (15)\\ &= \hat{N}(L)\hat{v}_t. && (16) \end{aligned}$$

What is important here is that $R_{\lambda_1}(L)\mathbf{K}'$ is a Blaschke matrix (Lippi & Reichlin 1994). Then, one can show that $\hat{v}_t$ is also orthonormal white noise such that the above model is indeed a VARMA model with polynomials $(A(L), \hat{N}(L))$. In order to satisfy the restriction $A_0 = M_0$,

we multiply by $C = \hat{N}_0^{-1} M_0$. Therefore, the invertible representation is

$$A(L)y_t = \hat{N}(L)CC^{-1}\hat{v}_t \quad (17)$$
$$A(L)y_t = \hat{M}(L)\hat{u}_t, \quad (18)$$

where $\Sigma_{\hat{u}} = C^{-1}(C^{-1})'$. If λ_1 is of multiplicity $m_1 > 1$ or if there are other roots with modulus smaller than one, then the described process has to be repeated until there are no more such roots. Finally, the residuals are calculated anew.

Also, if $(A(L), M(L))$ is in Echelon form, then $(A(L), \hat{M}(L))$ is in Echelon form with the same Kronecker indices. This can be seen as follow. $M(L)$ satisfies

$$m_{ki}(L) = \sum_{j=0}^{p_k} m_{ki,j} L^j \text{ with } M_0 = A_0 \ ,$$

for $k, i = 1, \ldots, K$. For $\hat{M}(L)$, we have $\hat{M}_0 = A_0$ by construction as outlined above. Furthermore,

$$\hat{M}(L) = M(L)P\mathbf{K}R_{\lambda_1}^{-1}(L)C$$

Therefore, the elements in $\hat{M}(L)$, $\hat{m}_{ki}(L)$, are linear combinations of elements of $M(L)$ with the *same* row index. That is, the above restrictions on the maximum lag order of polynomials in each row is satisfied.

B Figures and Tables

Table 1: Parameter values

DGP		Parameters	λ^{ar}	λ^{ma}
DGP I	MEV	$\alpha_{12,1} = 0.66,\ m_{21,1} = -0.13$	0.255 ± i 0.25 (0.36) 0.66	-0.05
	LPAREV	$\alpha_{12,1} = 0.90,\ m_{21,1} = -0.13$	0.255 ± i 0.25 (0.36) 0.9	-0.05
	LNAREV	$\alpha_{12,1} = -0.90,\ m_{21,1} = -0.13$	0.255 ± i 0.25 (0.36) -0.90	-0.05
	LPMAEV	$\alpha_{12,1} = 0.66,\ m_{21,1} = -2.25$	0.255 ± i 0.25 (0.36) 0.66	0.90
	LNMAEV	$\alpha_{12,1} = 0.66,\ m_{21,1} = 2.25$	0.255 ± i 0.25 (0.36) 0.66	-0.90
DGP II	MEV	$\alpha_{22,1} = 0.23,\ \alpha_{22,2} = 0.06$ $m_{22,1} = -0.75,\ m_{22,2} = 0.16$	0.39, -0.16	-0.38 ± i 0.14 (0.40)
	LPAREV	$\alpha_{22,1} = 0.744,\ \alpha_{22,2} = 0.14$ $m_{22,1} = -0.75,\ m_{22,2} = 0.16$	0.9, -0.16	-0.38 ± i 0.14 (0.40)
	LNAREV	$\alpha_{22,1} = -1.06,\ \alpha_{22,2} = -0.14$ $m_{22,1} = -0.75,\ m_{22,2} = 0.16$	-0.9, -0.16	-0.38 ± i 0.14 (0.40)
	LPMAEV	$\alpha_{22,1} = 0.23,\ \alpha_{22,2} = 0.06$ $m_{22,1} = -0.95,\ m_{22,2} = 0.25$	0.39, -0.16	-0.48 ± i 0.16 (0.50)
	LNMAEV	$\alpha_{22,1} = 0.23,\ \alpha_{22,2} = 0.06$ $m_{22,1} = 0.95,\ m_{22,2} = 0.25$	0.39, -0.16	0.48 ± i 0.16 (0.50)
DGP III	MEV	$\alpha_{11,1} = 0.7,\ m_{33,1} = 0.5$	0.7,0	0.5, -0.6
	LPAREV	$\alpha_{11,1} = 0.9,\ m_{33,1} = 0.5$	0.9,0	0.5, -0.6
	LNAREV	$\alpha_{11,1} = -0.9,\ m_{33,1} = 0.5$	-0.9,0	0.5, -0.6
	LPMAEV	$\alpha_{11,1} = 0.7,\ m_{33,1} = 0.9$	0.7,0	0.9, -0.6
	LNMAEV	$\alpha_{11,1} = 0.7,\ m_{33,1} = -0.9$	0.7,0	-0.9, -0.6
DGP IV	MEV	$\alpha_{11,1} = 0.5,\ m_{55,1} = -0.6$	0.5, 0 ± i 0.57 (0.57)	-0.6, -0.4 ± i 0.67 (0.78)
	LPAREV	$\alpha_{11,1} = 0.9,\ m_{55,1} = -0.6$	0.9, 0 ± i 0.57 (0.57)	-0.6, -0.4 ± i 0.67 (0.78)
	LNAREV	$\alpha_{11,1} = -0.9,\ m_{55,1} = -0.6$	-0.9, 0 ± i 0.57 (0.57)	-0.6, -0.4 ± i 0.67 (0.78)
	LPMAEV	$\alpha_{11,1} = 0.5,\ m_{55,1} = 0.9$	0.5, 0 ± i 0.57 (0.57)	0.9, -0.4 ± i 0.67 (0.78)
	LNMAEV	$\alpha_{11,1} = 0.5,\ m_{55,1} = -0.9$	0.5, 0 ± i 0.57 (0.57)	-0.9, -0.4 ± i 0.67 (0.78)

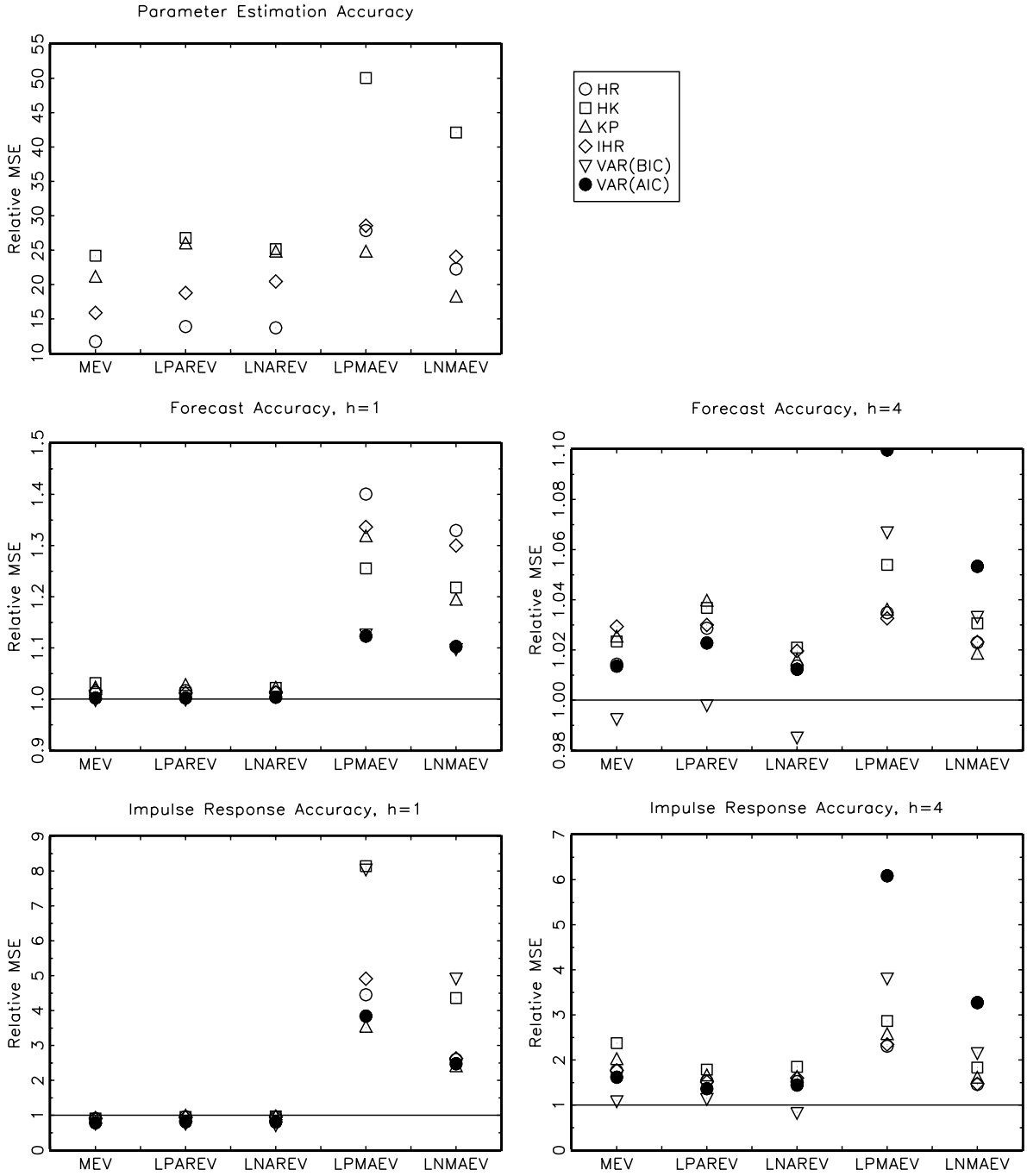
Note: Parameters and corresponding eigenvalues of the autoregressive and moving-average parts for the different data generating processes. The numbers in parenthesis denote moduli of complex eigenvalues.

Table 2: Estimation Failures, $T = 100$

DGP		HR	HK	K P	IHR
DGP I	MEV	2.5 / 0.0	17.9 / 0.5	0.0 / 0.0	1.1 / 0.1 /8.3
	LPAREV	2.5 / 0.0	16.7 / 0.9	0.1 / 0.0	1.2 / 0.0 /7.6
	LNAREV	2.5 / 0.0	19.5 / 0.7	0.1 / 0.0	1.4 / 0.0 /6.3
	LPMAEV	0.0 / 0.0	20.3 / 1.6	0.3 / 0.0	0.1 / 0.0 /76.2
	LNMAEV	0.0 / 0.0	14.2 / 0.8	0.2 / 0.0	0.0 / 0.0 / 81.5
DGP II	MEV	0.0 / 0.0	2.9 / 0.0	0.0 / 0.0	0.0 / 0.0 /14.8
	LPAREV	0.1 / 0.0	1.3 / 0.0	0.0 / 0.0	0.1 / 0.0 /16.5
	LNAREV	0.1 / 0.0	1.1 / 0.0	0.0 / 0.0	0.1 / 0.0 /13.1
	LPMAEV	0.0 / 0.0	3.8 / 0.0	0.0 / 0.0	0.0 / 0.0 /30.9
	LNMAEV	0.2 / 0.0	8.5 / 0.2	0.0 / 0.0	0.5 / 0.0 / 3.8
DGP III	MEV	0.0 / 0.0	0.9 / 0.0	0.1 / 0.1	0.0 / 0.0 /0.8
	LPAREV	0.0 / 0.0	1.2 / 0.0	0.1 / 0.0	0.0 / 0.0 /1.2
	LNAREV	0.0 / 0.0	0.7 / 0.0	0.1 / 0.0	0.0 / 0.0 /1.1
	LPMAEV	0.0 / 0.0	0.2 / 0.0	0.0 / 0.0	0.0 / 0.0 /3.8
	LNMAEV	0.0 / 0.0	0.2 / 0.0	0.1 / 0.0	0.0 / 0.0 / 4.1
DGP IV	MEV	0.0 / 0.0	1.6 / 0.0	0.1 / 0.0	0.0 / 0.0 /1.1
	LPAREV	0.1 / 0.0	0.4 / 0.0	0.6 / 0.0	0.1 / 0.0 /0.6
	LNAREV	0.0 / 0.0	0.4 / 0.0	0.2 / 0.0	0.1 / 0.0 /0.4
	LPMAEV	0.0 / 0.0	2.1 / 0.0	0.0 / 0.0	0.0 / 0.0 /2.4
	LNMAEV	0.0 / 0.0	1.3 / 0.0	0.2 / 0.0	0.1 / 0.0 / 2.6

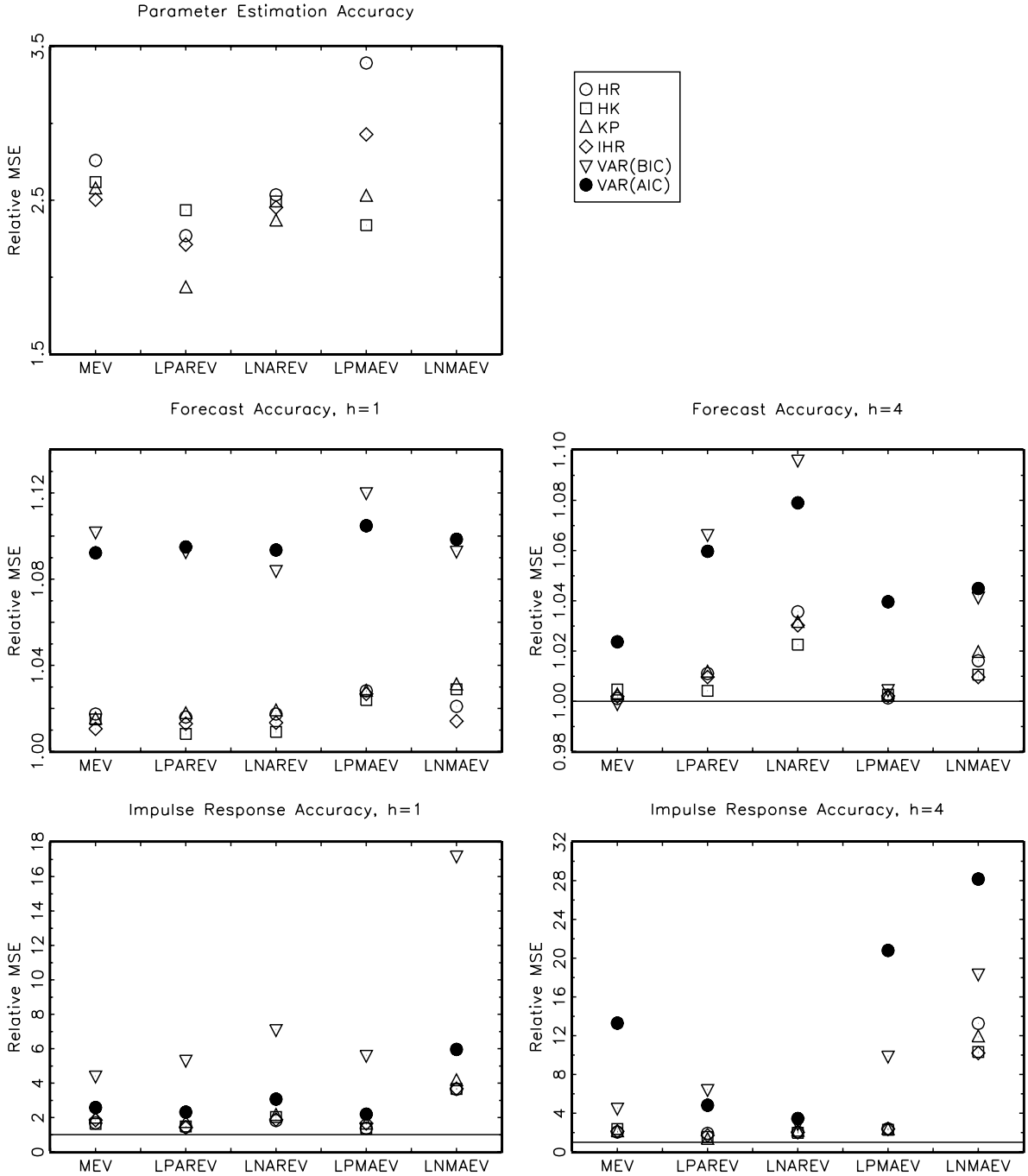
Note: Relative frequency of cases in which the algorithms returned non-stationary models / yielded extreme parameter values or, for IHR, did not converge (percentage points).

Figure 1: MSE ratios for DGP I with $T = 100$.



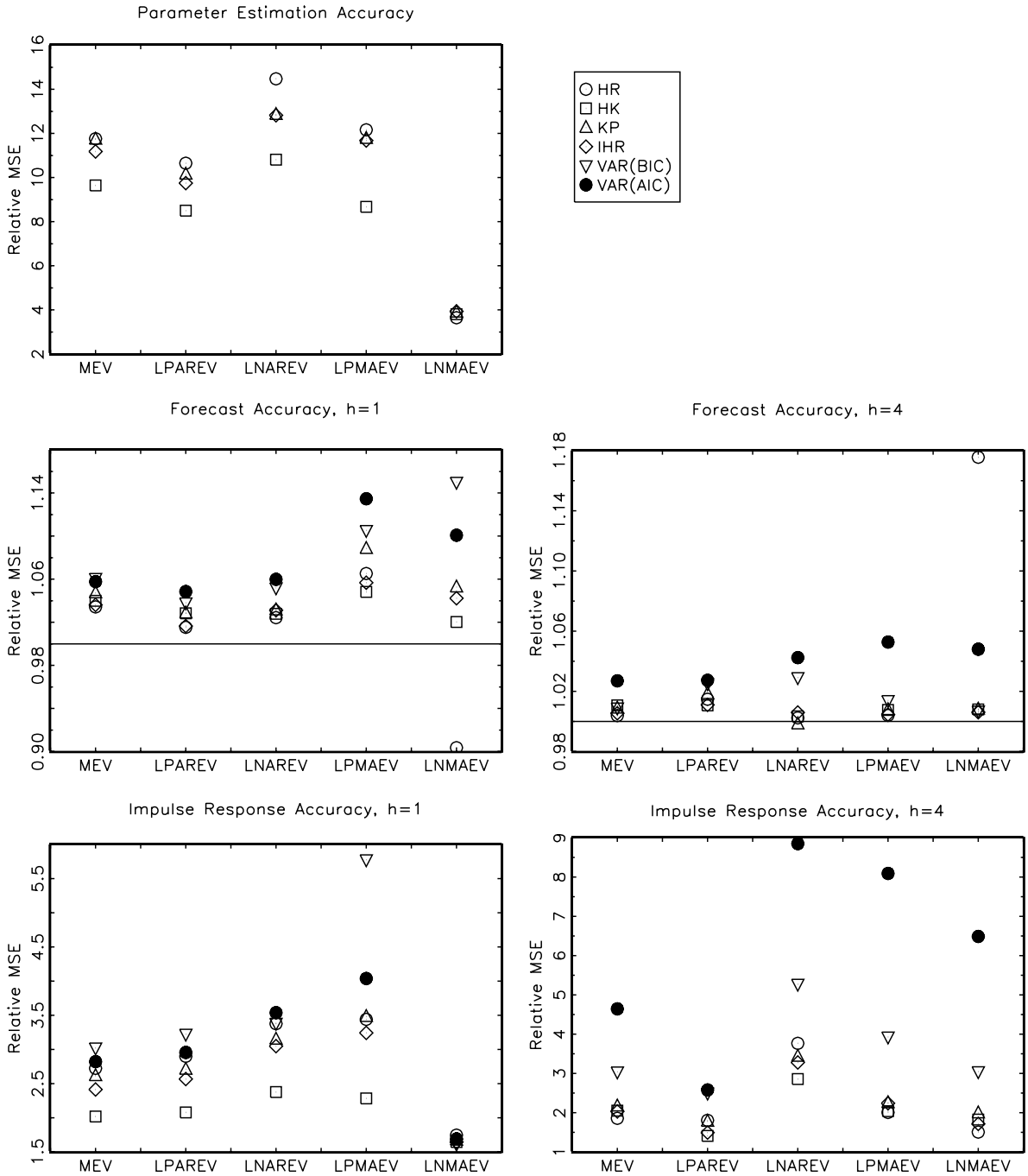
Note: On the vertical line, the graphs show relative mean squared error measures (section 4.1) related to the accuracy of the estimated vector of parameters, of predicted future values and of the estimated vector of impulse responses for different estimation algorithms (see legend). The benchmark algorithm is the maximum likelihood estimator. The horizontal line represents different parameterizations of the same model DGP.

Figure 2: MSE ratios for DGP II with $T = 100$.



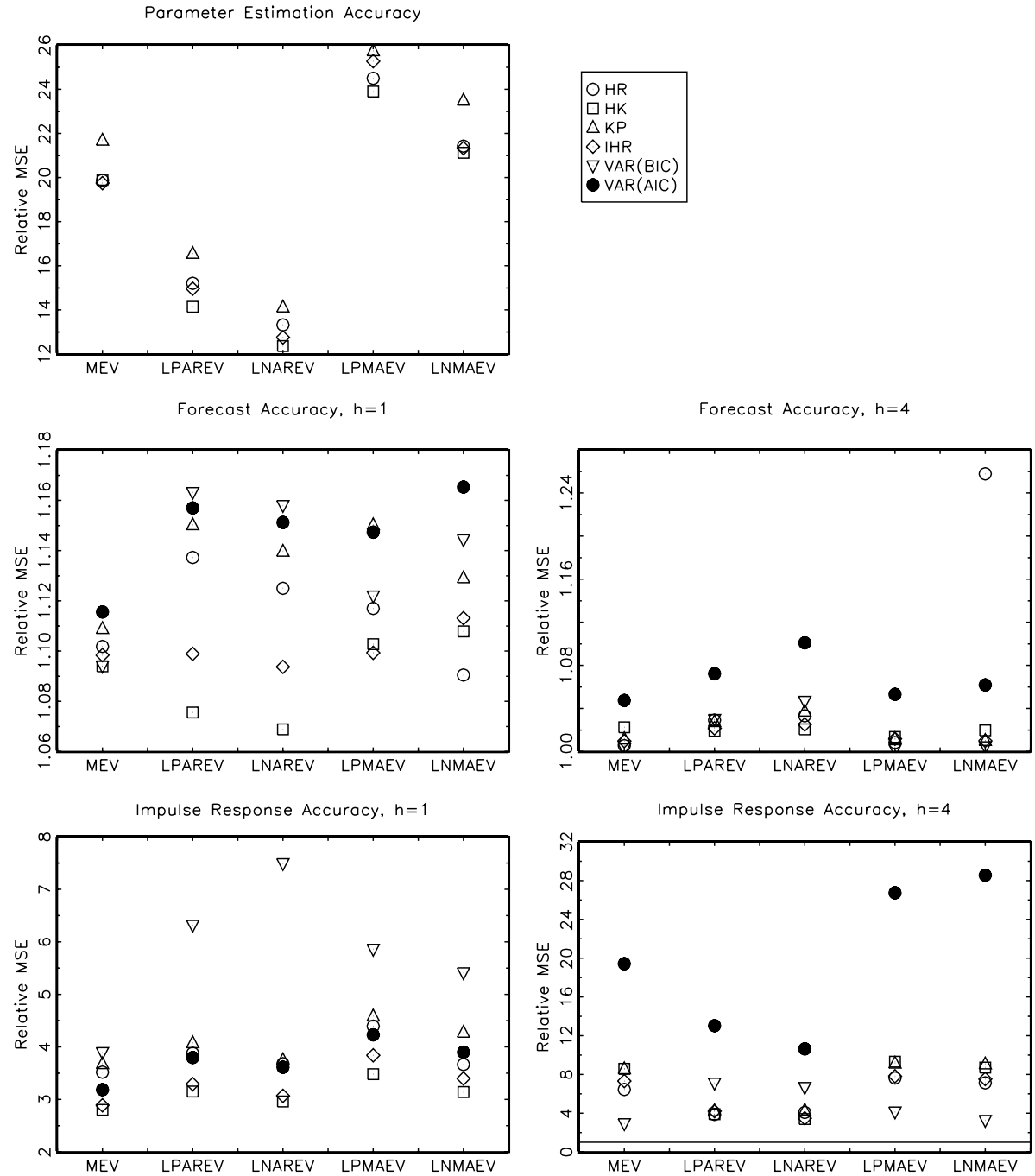
Note: See note to Figure 1.

Figure 3: MSE ratios for DGP III with $T = 100$.



Note: See note to Figure 1.

Figure 4: MSE ratios for DGP IV with $T = 100$.



Note: See note to Figure 1.