European University Institute

Department of Economics

EUI Working Papers

ECO 2007/12

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

Christian Kascha

EUROPEAN UNIVERSITY INSTITUTE
DEPARTMENT OF ECONOMICS

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

CHRISTIAN KASCHA

EUI Working Paper **ECO** 2007/12

ISSN 1725-6704

Printed in Italy
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy

http://www.eui.eu/
http://cadmus.eui.eu/

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

Christian Kascha†

March 2007

Abstract

Classical Gaussian maximum likelihood estimation of mixed vector autoregressive moving-average models is plagued with various numerical problems and has been considered difficult by many applied researchers. These disadvantages could have led to the dominant use of vector autoregressive models in macroeconomic research. Therefore, several other, simpler estimation methods have been proposed in the literature. In this paper these methods are compared 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.
†European University Institute, Department of Economics, Via della Piazzuola 43, 50133 Firenze, Italy; E-mail: christian.kascha@eui.eu.

1 Introduction

Although vector autoregressive moving-average (VARMA) models have theoretical and practical advantages compared to simpler vector autoregressive (VAR) models, VARMA models are rarely used in applied macroeconomic work. One likely reason is that the estimation of these models is considered difficult by many researchers. While Gaussian maximum likelihood estimation is theoretically attractive, it is plagued with various numerical problems. Therefore, simpler estimation algorithms have been proposed in the literature that, however, have not been compared systematically. In this paper some prominent estimation methods for VARMA models are compared by means of a Monte Carlo study. Different evaluation criteria such as the accuracy of point forecasts or the accuracy of the estimated impulse responses are used to judge the algorithms' performance. I focus on sample lengths and processes that could be considered typical for macroeconomic applications.

The problem of estimating VARMA models received a lot of attention for several reasons. While most economic relations are intrinsically nonlinear, linear models such as VARs or univariate autoregressive moving-average (ARMA) models have proved to be successful in many circumstances. They are 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). Therefore, VARMA models are of interest as generalizations of successful univariate ARMA models.

In the class of multivariate linear models, pure VARs are currently dominating in macroeconomic applications. These models have some drawbacks which could be overcome by the use of the more general class of VARMA models. First, 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. 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 (2005) show formally how DSGE models and VARMA processes are linked. Also Cooley & Dwyer (1998) claim that modelling macroeconomic time series systematically as pure VARs is not justified by the underlying economic theory. In sum, there are a number of theoretical reasons to prefer VARMA modelling to VAR modelling. However, there are also some complications that make VARMA modelling 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 parameter, the lag length. Thus, the search for an identified VARMA model is more complex than the specification of a VAR model. This aspect introduces additional uncertainty in the specification stage, although specification 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. The literature on estimation of VARMA models focussed on maximum likelihood methods which are asymptotically efficient (e.g. Hillmer & Tiao 1979, Mauricio 1995). However, the maximization of the Gaussian likelihood is not a trivial task. Numerical problems arise in the presence of nearly not-identified models, multiple equilibria and nearly non-invertible models. In high-dimensional, sparse systems maximum likelihood estimation may become even infeasible. In the specification stage one usually has do examine many different models which turn out not to be identified ex-post.

For these reasons several other estimation algorithms have been proposed in the literature. For example, Koreisha & Pukkila (1990) proposed a generalized least squares procedure. Kapetanios (2003) suggested an iterative least squares algorithm that uses only ordinary least

squares regressions at each iteration. 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 (1994) or Larimore (1983).[1] While there are nowadays several possible estimation methods available, it is not clear which methods are preferable under which circumstances. In this study some of these methods are compared by means of a Monte Carlo Study. 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. I conduct Monte Carlo simulations for four different DGPs with varying sample lengths and parameterizations. Five different simple algorithms are used and compared to maximum likelihood estimation and two benchmark VARs. The algorithms are a simple two-stage least squares algorithm, the iterative least squares procedure of Kapetanios (2003), the generalized least squares procedure of Koreisha & Pukkila (1990), a three-stage least squares procedure based on Hannan & Kavalieris (1984*a*) and the CCA subspace algorithm by Larimore (1983). The obtained results suggest that the algorithm of Hannan & Kavalieris (1984*a*) is the only algorithm that reliably outperforms 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 invertible. Therefore, the algorithm would have to be improved in order to make it an alternative tool for applied researchers.

The rest of the paper is organized as follows. In section 2 stationary VARMA processes and state space systems are introduced and some identified parameterizations are 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.

[1] See also the survey of Bauer (2005*b*).

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

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

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 K-dimensional 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 will also abstract from issues such as seasonality. As will be seen later, we consider models of the form (1) such that $A_0 = M_0$ and $A_0,\ M_0$ are non-singular. 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

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

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. 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 dimensions 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 approximate processes well that are actually generated by a VARMA system.

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 polynomials $A(L)$ and $M(L)$ 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 first possible source of non-uniqueness is that there are common factors in the polynomials that can be canceled out. For example, in a VARMA$(1,1)$ system such as

$$(I_K - A_1 L) y_t = (I_K + M_1 L) u_t$$

the autoregressive and the moving-average polynomial cancel out against each other if $A_1 = -M_1$. In order to ensure a unique representation we have to require that there are no common factors in both polynomials, that is $A(L)$ and $M(L)$ have to be *left-coprime*. This property may be defined by introducing the matrix operator $[A(L), M(L)]$ and calling it left-coprime if the existence of operators $D(L), \bar{A}(L)$ and $\bar{M}(L)$ satisfying

$$D(L)[\bar{A}(L), \bar{M}(L)] = [A(L), M(L)] \tag{5}$$

implies that $D(L)$ is unimodular.[2] A polynomial matrix $D(L)$ is called unimodular if its determinant, $|D(L)|$, is a nonzero constant that does not depend on L. Then $D(L)$ can only be of finite order having a finite order inverse. This condition ensures just that a representation is chosen for which further cancelation is not possible.

Still, the existence of many unimodular operators satisfying equation (5) cannot generally be ruled out. To obtain uniqueness of the autoregressive and moving-average polynomials we have to impose further restrictions ensuring that the only feasible operator satisfying the above equation is $D(L) = I_K$. Therefore, different representations have been proposed in the literature (Hannan & Deistler 1988, Lütkepohl 2005). These representations impose particular restrictions on the coefficient matrices that make sure that for a given process there is exactly one representation in the set of considered representations. We present two identified representations which are used later.

A VARMA(p, q) is in *final equations form* if it can be written as

$$\alpha(L)y_t = (I + M_1 + \ldots + M_q L^q)u_t,$$

where $\alpha(L) := 1 - \alpha_1 L - \ldots - \alpha_p L^p$ is a scalar operator with $\alpha_p \neq 0$. The moving-average polynomial is unrestricted apart from $M_0 = I_K$. It can be shown that this representation is uniquely identified provided that p is minimal (Lütkepohl 2005). A disadvantage of the final equations form is that it requires usually more parameters than other representations in order to represent the same stochastic process and thus might not be the most efficient representation.

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

[2] $[A, B]$ denotes a matrix composed horizontally of two matrices A and B.

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, $\alpha_{ki}(L),\ k \neq i$. Again, it can be shown that this representation leads to identified parameters (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 the Echelon Form with equal Kronecker indices, i.e. $p_1 = p_2 = \ldots = p_K$, corresponds to a standard unrestricted VARMA representation. This is one of the most promising representations, from a theoretical point of view, since it often leads to more parsimonious models than other representations.

There is also another representation of the same process which is algebraically equivalent. Every process that satisfies (1) can also be written as a state space model of the form

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t, \\
y_t &= Cx_t + u_t,
\end{aligned} \tag{6}$$

where the vector x_t is the so-called state vector of dimension $(n \times 1)$ and A $(n \times n)$, B $(n \times K)$, C $(K \times n)$ are fixed coefficient matrices. Generally, the state x_t is not observed. Processes that satisfy (6) can be shown to have a VARMA representation (see, e.g., Aoki 1989, Hannan & Deistler 1988). In the appendix it is illustrated how a VARMA model can be written in

state space form and how a state space model can define a VARMA model.

Also the state space representation is not identified unless restrictions on the parameter matrices are imposed. Analogously to the VARMA case, we first have to rule out over-parametrization by requiring that the order of the state vector, n, is minimal. Still, this does not determine a unique set (A, B, C) for a given process. To see this, consider multiplying the state vector by a nonsingular matrix $\mathbf{T}$ and define a new state vector $s_t := \mathbf{T}x_t$. The redefinition of the state leads to another state space representation given by

$$\begin{aligned} s_{t+1} &= \mathbf{T}A\mathbf{T}^{-1}s_t + \mathbf{T}Bu_t, \\ y_t &= C\mathbf{T}^{-1}s_t + u_t. \end{aligned}$$

Thus, the problem is to pin down a basis for the state x_t. There are various canonical parameterizations, among them parameterizations based on Echelon canonical forms. We briefly discuss here balanced canonical forms, in particular stochastic balancing, because it is used later in one of the estimation algorithms.

The discussion on stochastic balancing is based on Desai, Pal & Kirkpatrick (1985) and the introduction in Bauer (2005*a*). Define the observability matrix $\mathcal{O} := [C', A'C', (A^2)'C', \ldots]'$ and the matrix $\mathcal{K} := [B, (A-BC)B, (A-BC)^2B, \ldots]$. The unique parametrization is defined in terms of these matrices. Define as well an infinite vector of future observations as $Y_t^+ := (y_t', y_{t+1}', \ldots)'$ and an infinite vector of past observations as $Y_t^- := (y_{t-1}', y_{t-2}', \ldots)'$. Define analogously the vector of future residuals, U_t^+. Note that from (6) we can represent the state as a function of all past observations as

$$x_t = \mathcal{K}Y_t^-,$$

provided that the eigenvalues of $(A - BC)$ are less than one in modulus. The covariance matrix of the state vector is therefore given by $E[x_t x_t'] = \mathcal{K}E[Y_t^-(Y_t^-)']\mathcal{K}' = \mathcal{K}\Gamma_\infty^-\mathcal{K}'$, where $\Gamma_\infty^- := E[Y_t^-(Y_t^-)']$. Given a state space system as in (6), there is also another representation

which is called *backward innovation model*

$$\begin{aligned} z_t &= A' z_{t+1} + N f_t, \\ y_t &= M' z_{t+1} + f_t, \end{aligned}$$

where time "runs backwards" and A is as in (6) and N $(n \times K)$, M $(n \times K)$ are functions of (A, B, C), in particular $M = E[x_t y_{t-1}']$. The error f_t can be interpreted as the one-step ahead forecast error from predicting y_t given future observations. One can show that the variance of the backward state is given by $E[z_t z_t'] = \mathcal{O}'(E[Y_t^+(Y_t^+)'])^{-1}\mathcal{O} = \mathcal{O}'(\Gamma_\infty^+)^{-1}\mathcal{O}$, where $\Gamma_\infty^+ := E[Y_t^+(Y_t^+)']$.

Equipped with these definitions, we say that (A, B, C) and Σ is a stochastically balanced system if $E[x_t x_t'] = E[z_t z_t'] = diag(\sigma_1, \ldots, \sigma_n)$, with $1 > \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$. Also stochastically balanced systems are not unique. Uniqueness can however be obtained by determining the matrices $\mathcal{O}$ and $\mathcal{K}$ by means of the identification restrictions implicit in the singular value decomposition (SVD) for a given covariance sequence.

For doing so, introduce the *Hankel* matrix of autocovariances of y_t

$$\mathcal{H} := E[Y_t^+(Y_t^-)'] = \begin{bmatrix} \gamma(1) & \gamma(2) & \gamma(3) & \ldots \\ \gamma(2) & \gamma(3) & & \\ \gamma(3) & & & \\ \vdots & & & \end{bmatrix},$$

where $\gamma(j) := E[y_t y_{t-j}']$, $j = 1, 2, \ldots$ are the covariance matrices of the process y_t. Using the relation $\mathcal{H} = \mathcal{O}\mathcal{K}\Gamma_\infty^-$, a stochastically balanced representation can be obtained by using the SVD of

$$(\Gamma_\infty^+)^{-1/2}\mathcal{H}\left[(\Gamma_\infty^-)^{-1/2}\right]' = U_n S_n V_n'.$$

Setting $\mathcal{O} = (\Gamma_\infty^+)^{1/2} U_n S_n^{1/2}$ and $\mathcal{K} = S_n^{1/2} V_n' (\Gamma_\infty^-)^{-1/2}$, the associated system is in balanced

form with $E[x_t x_t'] = E[z_t z_t'] = S_n$.[3] From the definition of the parametrization one can see that it is not easy to incorporate prior knowledge of parameter restrictions.

While the VARMA and the state space representation are equivalent in an algebraic sense, they lead to other estimation techniques and therefore differ in a statistical sense. These models have become popular because of their conceptional simplicity and because they allow for estimation algorithms, namely so-called subspace methods, that possess very good numerical properties. See also Deistler, Peternell & Scherrer (1995) and Bauer (2005*b*) for some of the properties of subspace algorithms. It is claimed that these methods are very successful in estimating multivariate linear systems. Therefore, subspace methods are also considered as potential competitors to estimation techniques which rely on the more standard VARMA representation.

3 Description of Estimation Methods

In the following, a short description of the examined algorithms is given. Obviously, one cannot consider each and every existing algorithm but only a few popular algorithms. The hope is that the performance of these algorithms indicate how their variants would work. Throughout it is assumed that the data has been mean-adjusted prior to estimation. In the following, 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 throughout it is assumed that restrictions are imposed on the parameter vector of the VARMA model. That is, the coefficient matrices are assumed to be restricted according to the final equations form or the Echelon form. I adopt the following notation. The observed sample is $y_1, y_2, \ldots, y_T$. I denote the vector of total parameters by β ($K^2(p+q)\times 1$) and the vector of free parameters by γ. Let the dimension of γ be given by n_γ. 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

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

[3] The square root of a matrix X, $Y = X^{1/2}$ is defined such that $YY' = X$.

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. A_0 is assumed to be either the identity matrix or to satisfy the restrictions imposed by the Echelon representation. To consider zero and equality restrictions on the parameters, define a $((K^2(p+q)) \times n_\gamma)$ matrix R such that

$$\beta = R\gamma. \tag{7}$$

This notation is equivalent to the explicit formulation of restrictions on β such as $C\beta = c$ for suitable matrices C and c. The above notation, however, is advantageous for the representation of the estimation algorithms.

Two-Stage Least Squares (2SLS) This is the simplest method. 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

$$y_t = \sum_{i=1}^{n_T} \Pi_i y_{t-i} + u_t, \tag{8}$$

where n_T is the lag length that is required to increase with the sample size, T. In the second stage, the residuals from (8), $\hat{u}_t^{(0)}$, are plugged in (1). After rearranging (1), one gets

$$\begin{aligned} y_t &= (I_K - A_0)[y_t - \hat{u}_t^{(0)}] + A_1 y_{t-1} + \ldots + A_p y_{t-p} \\ &\quad + M_1 \hat{u}_{t-1}^{(0)} + \ldots + M_q \hat{u}_{t-q}^{(0)} + u_t, \end{aligned} \tag{9}$$

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

$$Y_{t-1}^{(0)} := \begin{bmatrix} (y_t - \hat{u}_t^{(0)}) \\ y_{t-1} \\ \vdots \\ y_{t-p} \\ \hat{u}_{t-1}^{(0)} \\ \vdots \\ \hat{u}_{t-q}^{(0)} \end{bmatrix}.$$

Collecting all observations we get

$$Y = [I_K - A_0, \mathbf{A}, \mathbf{M}]X^{(0)} + U, \tag{10}$$

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 (10) yields

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

and the 2SLS estimator is defined as

$$\tilde{\gamma} = [R'(X^{(0)}X^{(0)\prime} \otimes \widetilde{\Sigma}^{-1})R]^{-1}R'(X^{(0)} \otimes \widetilde{\Sigma}^{-1})\text{vec}(Y). \tag{11}$$

where $\widetilde{\Sigma}$ is the covariance matrix estimator based on the residuals $\hat{u}_t^{(0)}$. 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. Alternatively, one may also plug in the estimated current innovation $\hat{u}_t^{(0)}$ in (9), define a new regression error, say ξ_t, and regress $y_t - \hat{u}_t^{(0)}$ on $Y_{t-1}^{(0)}$. Existing Monte Carlo studies though indicate that the difference between both variants is of minor importance (Koreisha & Pukkila 1989).

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). Lütkepohl & Poskitt (1996) propose for multivariate, non-seasonal data a value between $\log T$ and $\sqrt{T}$. Throughout the whole paper we employ $n_T = 0.5\sqrt{T}$.[4]

Hannan-Kavalieris-Procedure (3SLS) This method adds a third stage to the procedure just described. It goes originally back to Durbin (1960) and has been introduced by Hannan & Kavalieris (1984*a*) for multivariate processes.[5] 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 been sometimes interpreted as a three-stage least squares procedure (Dufour & Pelletier (2004)). The method is computationally very easy to implement because of its recursive nature. Corresponding to the estimates of the 2SLS algorithm, new residuals, ε_t $(K \times 1)$, are formed. One step of the Gauss-Newton iteration is performed starting from these estimates. For this reason, matrices, ξ_t $(K \times 1)$, η_t $(K \times 1)$ and $\hat{X}_t$ $(K \times n_\gamma)$ are calculated according to

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

[4]Since this algorithm provides also starting values for other algorithms, it is quite important that the resulting estimated VARMA model is invertible. In case the initial estimate does not imply an invertible VARMA model, different lag lengths are tried in order to obtain an invertible model. If this procedure fails, the estimated moving-average polynomial, say $\widehat{M}(L)$, is replace by $\widehat{M}_\lambda(L) = \hat{M}_0 + \lambda(\widehat{M}(L) - \hat{M}_0),\ \lambda \in (0,1)$. The latter case occurs in less than 0.1 % of the cases.

[5]See also Hannan & Deistler (1988), sections 6.5, 6.7, for an extensive discussion.

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 3SLS 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. While the 2SLS estimator is not asymptotically efficient, the 3SLS is, because it performs one iteration of a conditional maximum likelihood procedure starting from the estimates of the 2SLS procedure.

Hannan & Kavalieris (1984*b*) showed consistency and asymptotic normality of these estimators. Dufour & Pelletier (2004) extend these results to even more general conditions. The Monte Carlo evidence presented by Dufour & Pelletier (2004) indicates that this estimator represents a good alternative to maximum likelihood in finite samples. It is possible to use this procedure iteratively, starting the above recursions in the second iteration with the newly obtained parameter estimates in $\hat{\gamma}$ from the 3SLS procedure, and so on until convergence.

Generalized Least Squares (GLS) Also this procedure has three stages. Koreisha & Pukkila (1990a) proposed this procedure for univariate ARMA models and Kavalieris, Hannan & Salau (2003) proved efficiency of the GLS estimates in this case. See also Flores de Frutos & Serrano (2002). The motivation is the same as for the 2SLS estimator. Given consistent estimates of the residuals, we can estimate the parameters of the VARMA representation by least squares. However, Koreisha & Pukkila (1990a) 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 GLS procedure tries to take this into account. I consider a multivariate generalization of the same procedure. In the first stage, preliminary estimates of the innovations are obtained by a long autoregression as in (8). Koreisha & Pukkila (1990a) *assume* that the residuals obtained from (8) estimate 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

$$\begin{aligned} A_0 y_t &= \sum_{j=1}^{p} A_j y_{t-j} + A_0(\hat{u}_t^{(0)} + \epsilon_t) + \sum_{j=1}^{q} M_j(\hat{u}_{t-j}^{(0)} + \epsilon_{t-j}), \\ y_t &= (I_K - A_0)(y_t - \hat{u}_t^{(0)}) + \sum_{j=1}^{p} A_j y_{t-j} + \hat{u}_t^{(0)} \\ &\quad + \sum_{j=1}^{q} M_j \hat{u}_{t-j}^{(0)} + A_0\epsilon_t + \sum_{j=1}^{q} M_j \epsilon_{t-j}, \\ y_t - \hat{u}_t^{(0)} &= (I - A_0)(y_t - \hat{u}_t^{(0)}) + \sum_{j=1}^{p} A_j y_{t-j} \\ &\quad + \sum_{j=1}^{q} M_j \hat{u}_{t-j}^{(0)} + \zeta_t. \end{aligned} \tag{12}$$

As can be seen from these equations, the error term, ζ_t, in a regression of y_t on its lagged values and estimated residuals $\hat{u}_t^{(0)}$ is not uncorrelated but is a moving-average process of order q, $\zeta_t = A_0\epsilon_t + \sum_{j=1}^{q} M_j\epsilon_{t-j} = \tilde{\epsilon}_t + \sum_{j=1}^{q} M_j A_0^{-1}\tilde{\epsilon}_{t-j}$, where $\tilde{\epsilon}_t := A_0\epsilon_t$. Thus, a least squares regression in (12) is not efficient. Koreisha & Pukkila (1990a) propose the following three-stage algorithm to take the correlation structure of ζ_t into account. In the first stage the residuals are estimated using a long autoregression. In the second stage one estimates the coefficients in (12) 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 2SLS 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, $\Sigma_\zeta := E[\zeta_t\zeta_t']$, is estimated as $\tilde{\Sigma}_\zeta = T^{-1}\sum \tilde{\tilde{\zeta}}_t(\tilde{\tilde{\zeta}}_t)'$. From the relations $\zeta_t = A_0\epsilon_t + M_1\epsilon_{t-1} + \ldots + M_q\epsilon_{t-q}$ and $\Sigma_\zeta = A_0\Sigma_\epsilon A_0' + \ldots + M_q\Sigma_\epsilon M_q'$ one

can retrieve

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

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 (12) by GLS using $\hat{\Phi}$:

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

In comparison to the 2SLS estimator the main difference lies in the GLS weighting with $\hat{\Phi}^{-1}$. Given $\hat{\tilde{\gamma}}$ one could calculate new estimates of the residuals ζ_t and update the estimate of the covariance matrix. Given these quantities one would obtain a new estimate of the parameter vector and so on until convergence.[6]

Iterative Least Squares (IOLS) The suggestion made by Kapetanios (2003) is simply to use the 2SLS algorithm iteratively. Denote the estimate of the 2SLS procedure by $\tilde{\gamma}^{(1)}$. We may obtain new residuals by

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

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 (11) in

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

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 $||\hat{U}^{(i-1)} - \hat{U}^{(i)}|| < c$ according to some

[6]The evidence given by Koreisha & Pukkila (1990a), however, suggests that further iterations do have a negligible effect. This is also the experience of the present author. The results presented here are therefore given for the first iteration of the GLS procedure.

pre-specified number c. In contrast to the above-mentioned regression-based procedures, the IOLS procedure is iterative but the computational load is still minimal.

Maximum Likelihood Estimation (MLE) The dominant approach to the estimation of VARMA models has been of course maximum likelihood estimation. Given a sample, $y_1, ..., y_T$, the Gaussian likelihood conditional on initial values can be easily set up as

$$l(\gamma) = \sum_{t=1}^{T} l_t(\gamma)$$

where

$$\begin{aligned} l_t(\gamma) &= -\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}u_t'(\gamma)\Sigma^{-1}u_t(\gamma), \\ u_t(\gamma) &= M_0^{-1}\big(A_0 y_t - A_1 y_{t-1} - \ldots - A_p y_{t-p} \\ &\quad -M_1 u_{t-1}(\gamma) - \ldots - M_q u_{t-q}(\gamma)\big). \end{aligned}$$

The initial values for y_t and u_t are assumed to be fixed equal to zero (see Lütkepohl 2005). These assumptions introduce a negligible bias if the orders of the VARMA model are low and the roots of the moving-average polynomial are not close to the unit circle. In contrast, exact maximum likelihood estimation does consider the exact, unconditional likelihood that backcasts the initial values. The formulation of this procedure requires some considerable investment in notation and can be found for example in Reinsel (1993). Since processes with large moving-average eigenvalues are also investigated, exact maximum likelihood estimation is considered. The procedure is implemented using the time series package 4.0 in GAUSS. The algorithm is based upon the formulation of Mauricio (1995) and uses a modified Newton-algorithm. The starting values are the true parameter values and therefore the results from the exact maximum likelihood procedure must be regarded as a benchmark than as a realistic estimation alternative.

Subspace Algorithms (CCA) Subspace algorithms rely on the state space representation of a linear system. There are many ways to estimate a state space model, e.g., Kalman-

based maximum likelihood methods and subspace identification methods such as N4SID of Van Overschee & DeMoor (1994) or the CCA method of Larimore (1983). In addition, many variants of the standard subspace algorithms have been proposed in the literature. I focus only on one subspace algorithm, the CCA algorithm. The algorithm is asymptotically equivalent to maximum likelihood and was previously found to be remarkably accurate in small samples and is likely to be well suited for econometric applications (see Bauer 2005*b*). The general motivation for the use of subspace algorithms lies in the fact that if we knew the unobserved state, x_t, we could estimate the *system matrices*, A, B, C, by linear regressions as can be seen from the basic equations

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t \\ y_t &= Cx_t + u_t. \end{aligned}$$

Given knowledge of the state, estimates, $\hat{C}$ and $\hat{u}_t$, could be obtained by a regression of y_t on x_t and $\hat{A}$ and $\hat{B}$ could be obtained by a regression of x_{t+1} on x_t and $\hat{u}_t$. Therefore, one obtains in a first step an estimate of the n-dimensional state, $\hat{x}_t$. This is analogous to the idea of a long autoregression in VARMA models that estimates the residuals in a first step that is followed by a least squares regression. Solving the state space equations, one can express the state as a function of past observations of y_t and an initial state for some integer $\mathfrak{p} > 0$ as

$$\begin{aligned} x_t &= (A - BC)^{\mathfrak{p}} x_{t-\mathfrak{p}} + \sum_{i=0}^{\mathfrak{p}-1} (A - BC)^i B y_{t-i-1}, \\ &= (A - BC)^{\mathfrak{p}} x_{t-\mathfrak{p}} + \mathcal{K}_{\mathfrak{p}} Y_{t,\mathfrak{p}}^-, \end{aligned} \tag{13}$$

where $\mathcal{K}_{\mathfrak{p}} = [B, (A - BC)B, \ldots, (A - BC)^{\mathfrak{p}-1}B]$ and $Y_{t,\mathfrak{p}}^- = [y'_{t-1}, \ldots, y'_{t-p}]'$. On the other hand, one can express future observations as a function of the current state and future noise as

$$y_{t+j} = CA^j x_t + \sum_{i=0}^{j-1} CA^i B u_{t+j-i-1} + u_{t+j}. \tag{14}$$

Therefore, at each t, the best predictor of y_{t+j} is a function of the current state only, $CA^j x_t$, and thus the state summarizes in a certain sense all available information in the past up to time t.

Define $Y_{t,f}^+ = [y_t', \ldots, y_{t+f-1}']'$ for some integer $f > 0$ and formulate equation (14) for all observations contained in $Y_{t,f}^+$ simultaneously. Combine these equations with (13) in order to obtain

$$Y_{t,f}^+ \quad = \quad \mathcal{O}_f \mathcal{K}_{\mathfrak{p}} Y_{t,\mathfrak{p}}^- + \mathcal{O}_f (A - BC)^{\mathfrak{p}} x_{t-\mathfrak{p}} + \mathcal{E}_f E_{t,f}^+$$

where $\mathcal{O}_f = [C', A'C', \ldots, (A^{f-1})'C']'$, $E_{t,f}^+ = [u_t', \ldots, u_{t+f-1}']'$ and $\mathcal{E}_f$ is a function of the system matrices. The above equation is central for most subspace algorithms. Note that if the maximum eigenvalue of $(A-BC)$ is less than one in absolute value we have $(A-BC)^{\mathfrak{p}} \approx 0$ for large $\mathfrak{p}$. This condition is called the *minimum phase assumption*. This reasoning motivates an approximation of the above equation given by

$$Y_{t,f}^+ \quad = \quad \beta Y_{t,\mathfrak{p}}^- + N_{t,f}^+ \tag{15}$$

where $\beta = \mathcal{O}_f \mathcal{K}_{\mathfrak{p}}$ and $N_{t,f}^+$ is defined by the equation. Most popular subspace algorithms use this equation to obtain an estimate of β which is decomposed into $\mathcal{O}_f$ and $\mathcal{K}_{\mathfrak{p}}$. The identification problem is solved implicitly during this step. Different algorithms use these matrices differently to obtain an estimate of the state. Given an estimate of the state, the system matrices are recovered.

For given integers $n, \mathfrak{p}, f$, the employed algorithm consists of the following steps :

1. Set up $Y_{t,f}^+$ and $Y_{t,\mathfrak{p}}^-$ and perform OLS in (15) using the available data to get an estimate $\hat{\beta}_{f,\mathfrak{p}}$.

2. Compute the sample covariances

$$\hat{\Gamma}_f^+ = \frac{1}{T_{f,\mathfrak{p}}} \sum_{t=\mathfrak{p}+1}^{T-f+1} Y_{t,f}^+ (Y_{t,f}^+)' \ , \ \hat{\Gamma}_{\mathfrak{p}}^- = \frac{1}{T_{f,\mathfrak{p}}} \sum_{t=\mathfrak{p}+1}^{T-f+1} Y_{t,\mathfrak{p}}^- (Y_{t,\mathfrak{p}}^-)',$$

where $T_{f,\mathfrak{p}} = T - f - \mathfrak{p} + 1$.

3. Given the dimension of the state, n, compute the singular value decomposition

$$(\hat{\Gamma}_f^+)^{-1/2}\hat{\beta}_{f,\mathfrak{p}}(\hat{\Gamma}_{\mathfrak{p}}^-)^{1/2} = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}_n,$$

where $\hat{\Sigma}_n$ is a diagonal matrix that contains the n largest singular values and $\hat{U}_n$ and $\hat{V}_n$ are the corresponding singular vectors. The remaining singular values are neglected and the approximation error is $\hat{R}_n$. The reduced rank matrices are obtained as

$$\begin{aligned}\hat{\mathcal{O}}_f &= [(\hat{\Gamma}_f^+)^{1/2}\hat{U}_n\hat{\Sigma}_n^{1/2}],\\ \hat{\mathcal{K}}_{\mathfrak{p}} &= [\hat{\Sigma}_n^{1/2}\hat{V}_n'(\hat{\Gamma}_{\mathfrak{p}}^-)^{-1/2}].\end{aligned}$$

4. Estimate the state as $\hat{x}_t = \hat{\mathcal{K}}_{\mathfrak{p}}Y_{t,\mathfrak{p}}^-$ and estimate the system matrices using linear regressions as described above.

Although the algorithm looks quite complicated at first sight, it is actually very simple and is regarded to lead to numerically stable and accurate estimates. There are certain parameters which have to be determined before estimation. While the order of the system is given by the simulated process, the integers $f, \mathfrak{p}$ have to be chosen deterministically or data-dependent. For example, Deistler et al. (1995) advocated choosing $f = \mathfrak{p} = dp_{BIC}$ for some $d > 1$, while in the paper of Bauer (2005*a*) $f = \mathfrak{p} = 2p_{AIC}$ is suggested, where p_{BIC} and p_{AIC} are the orders chosen by the BIC and AIC criterion for an autoregressive approximation, respectively. Here $f = \mathfrak{p} = 2p_{AIC}$ is employed.

4 Monte Carlo Study

I compare the performance of the different estimation methods using a variety of measures that could reveal possible gains of VARMA modelling. Namely, 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, it is also known that simple univariate models such as an AR(1) can outperform much more general models or even the correct model in terms of forecasting precision. This phenomenon is simply due to the limited information in small samples. Analogously, algorithms that may be asymptotically sub-optimal, may still be preferable when it comes to forecasting in small samples. With the sample size tending to infinity, the more exact algorithms will also yield better forecasts, but this might not be true for the small sample sizes investigated. 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. The most relevant statistic is the number of cases when the algorithms yielded non-invertible VARMA models. In this case the resulting residuals cannot be interpreted as prediction errors anymore. For the IOLS algorithm another relevant statistic is the number of cases when the iterations did not converge. These statistics are defined more precisely in section 4.3. In both cases and for all algorithms the estimates of the 2SLS procedure are adopted as the result of the particular algorithm for the corresponding replication of the simulation experiment.

I consider various processes and variations of them as described below. For all data generating processes I simulate $N = 1000$ series of length $T = 100$ and $T = 200$. The index n refers to a particular replication of the simulation experiment. The sample sizes represent typical lengths of data in macroeconomic time series applications. The investigated processes include small-dimensional and higher-dimensional systems. I consider mostly processes that have been used in the literature to demonstrate the virtue of specific algorithms but I also consider an example taken from estimated processes.

4.1 Performance Measures

4.1.1 Parameter Estimates

The accuracy of the different parameter estimates 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. One would like to summarize the accuracy of an estimator by a weighted average of its squared deviations from the true value. That is, for each algorithm the following statistic is computed

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

Here, Σ_{γ} denotes the large sample variance of the parameter estimates obtained by exact maximum likelihood. In order to ease interpretation, we compute the ratio of the MSE of a particular algorithm relative to the mean squared error of the MLE method:

$$\frac{MSE_{\mathcal{A}}}{MSE_{MLE}}.$$

4.1.2 Forecasting

Forecasting is one of the main objectives in time series modelling. To assess the forecasting power of different VARMA estimation algorithms I compare forecast mean squared errors (FMSE) of 1-step and 4-step ahead out-of-sample forecasts. I calculate the FMSE at horizon h for the algorithm $\mathcal{A}$ as

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

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, where the dependence on $\mathcal{A}$ is suppressed. The covariance matrix Σ_h refers to the corresponding theoretical h-step ahead forecast error obtained by using the true model with known parameters based on the information set $\Omega_T =$

$\{y_s|s \leq T\}$, that is, on all past data. Then the forecast MSE matrix turns out to be

$$\Sigma_h = \sum_{i=0}^{h-1} \Phi_i \Sigma \Phi_i'.$$

For given estimated parameters and a finite sample at hand, the white noise sequence u_t can be estimated recursively, using the past data as $u_t = y_t - A_0^{-1}\left(\sum_{i=1}^{p} A_j y_{t-j} + \sum_{j=1}^{q} M_j u_{t-j}\right)$, given some appropriate starting values, $u_0, u_{-1}, \ldots, u_{-q+1}$ and $y_0, y_{-1}, \ldots, y_{-p+1}$. These are computed using the algorithm of Mauricio (1995). The obtained residuals, $\hat{u}_t$, are used to compute the 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 unrestricted long VAR approximation:

$$\frac{FMSE_{\mathcal{A}}(h)}{FMSE_{\text{VAR}}(h)}.$$

In addition, I also compute the FMSE of a standard unrestricted VAR with lag length chosen by the AIC criterion in order to assess the potential merits of VARMA modelling compared to standard VAR modelling.

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. To assess the accuracy of the estimated impulse response function 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

$$IRMSE(h) = \frac{1}{N}\sum_{n=1}^{N}(\psi_h - \hat{\psi}_{h,n})'\Sigma_{\psi,h}^{-1}(\psi_h - \hat{\psi}_{h,n}),$$

where ψ_h is the theoretical response of y_{t+h} to shocks in u_t and $\hat{\psi}_{h,n}$ is the estimated response. $\Sigma_{\psi,h}$ is the asymptotic variance-covariance matrix of the impulse response function estimates obtained by maximum likelihood estimation. The precision of the estimated responses are again measured relative to the long VAR:

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

Also in this case, the results for a VAR with lag length chosen by the AIC criterion are computed.

4.2 Generated Systems

4.2.1 Small-Dimensional Systems

DGP I: The first two-dimensional process has been taken from Kapetanios (2003). This is a simple bivariate system in final equations form and was used in Kapetanios's (2003) paper to demonstrate the virtues of the IOLS procedure. Precisely, the process is given by

$$y_t = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_1 \end{pmatrix} y_{t-1} + u_t + \begin{pmatrix} m_{11,1} & -0.20 \\ 0.15 & m_{22,1} \end{pmatrix} u_{t-1}$$

and

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

This is an admittedly very simple process that is supposed to give an advantage to the IOLS procedure and also serves as a best case scenario for the VARMA algorithms because of its simplicity.

The autoregressive polynomial has one eigenvalue and the moving-average polynomial has two distinct eigenvalues 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 remaining parameters, α_1, $m_{11,1}$ and $m_{22,1}$ 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 the original process used in Kapetanios's (2003) paper, with $\alpha_1 = 0.2$, $m_{11,1} = 0.25$ and $m_{22,1} = -0.10$. I fit restricted VARMA models in final equations form to the data. This gives a slight advantage to algorithms based on the VARMA formulation since in this case the CCA method has to

estimated relatively more parameters. For the CCA method the dimension of the state vector is set to the true McMillian degree which is two.

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 $\lambda_1^{ar} = 0.385$ $\lambda_2^{ar} = -0.159$, $\lambda_1^{ma} = 0.375 + 0.139i$, $\lambda_2^{ma} = 0.375 - 0.139i$. Restricted 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 illus-

trative 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 feedback mechanisms are complex. The data is generated according to

$$y_t = \begin{pmatrix} \alpha_{11,1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0.4 & 0 \end{pmatrix} y_{t-1} + u_t + \begin{pmatrix} 0 & 1.1 & 0 \\ 0 & m_{22,1} & 0 \\ 0 & 0 & 0.5 \end{pmatrix} u_{t-1}$$

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 models are fit to the data. While this DGP is of higher dimension, the associated parameter matrices are more sparse. This property is reflected in the fact that the autoregressive polynomial and the moving-average polynomial have both only one root different from zero.

The parameters $\alpha_{11,1}$ and $m_{22,1}$ 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 and is thought to typify many practical real data applications. In this study 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 fit 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 $\lambda_1^{ar} = 0.5,\ \lambda_2^{ar} = 0 \pm i0.57,\ \lambda_1^{ma} = -0.6$ and $\lambda_2^{ma} = -0.4 \pm i0.67$.

4.3 Results

The results are summarized in tables 2 to 5 and figures 1 to 8. The tables show the frequency of cases when the algorithms failed for different reasons. The figures plot the various MSE ratios discussed above. In the tables and figures, 2SLS, 3SLS and GLS are the two-stage, three-stage and generalized least squares methods, respectively, IOLS is the iterative least squares algorithm, CCA denotes the CCA subspace algorithm, SVAR is the VAR chosen by AIC and MLE is the maximum likelihood algorithm.

Table 2 and table 3 display the frequency of cases when the algorithms yielded models that were not invertible or, in the case of the CCA algorithm, violated the minimum phase assumption for sample sizes $T = 100$ and $T = 200$, respectively. Apart from these cases,

there are also cases when the IOLS algorithm did not converge. The IOLS algorithm is regarded as non-convergent if it did not converge after 500 iterations. Furthermore, there are some very rare instances when the GLS algorithm returned estimated models that were extremely far from the true process (0.8 % in DGP III, LNMAEV, T=200). The tables 4 and 5 show the frequency of cases when the algorithms failed for one of the mentioned reasons in order to give a comprehensive picture of the reliability of the algorithms. First, as expected, the algorithms yield non-invertible models more frequently when the eigenvalues of the moving-average polynomial are close to one in absolute value, in particular in the case of large negative eigenvalues. Furthermore, as the number of estimated parameters increases, the algorithms yield non-invertible models more often. For 200 observations all algorithms become much more reliable in the sense that the number of estimated non-invertible models is much reduced. The most reliable algorithms are 2SLS, GLS and CCA. In particular, GLS and CCA yield non-invertible models for all algorithms and sample sizes in less than 1% of the replications. The 3SLS and the IOLS algorithm are the less reliable algorithms, although the IOLS algorithm can be quite stable. For particular DGPs, 3SLS can occasionally yield non-invertible models in more than 10 % of the cases. For some DGPs the IOLS algorithm does not converge relative frequently, but the problem becomes much less severe when the number of observations is increased to $T = 200$ as can be seen from tables 4 and 5.

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 differences become also more pronounced when the number of observations increases but the ranking of the algorithms remains unchanged, in general. The 2SLS algorithm is dominated by the other algorithms, aside from one case (DGP III, LNMAEV). The parameter estimation accuracy of the GLS estimator is much better but close to the accuracy of the 2SLS algorithm for higher-order processes, although in cases with large negative moving-average eigenvalues the estimator might be relatively accurate. The IOLS estimator is in most cases much better than 2SLS and its advantage becomes most pronounced in the high-dimensional case IV. Compared to the GLS algorithm, IOLS can be worse for small-dimensional systems but the ranking changes for the higher

dimensional processes. The 3SLS estimator is always superior to any other method, apart from MLE. In particular, the 3SLS method is much better when the number of estimated parameters increases, that is for DGP III and IV. However, the MLE method is in this context much more accurate and is often twice as good as the 3SLS method. Summarizing, the 3SLS procedure is the best alternative to MLE despite the high number of cases when the algorithm yielded non-invertible model. Nevertheless, even the best alternative can be quite imprecise compared to MLE. This does not necessarily mean that 3SLS 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 do estimate the parameters more accurately than others, they are not necessarily superior in terms of forecasting accuracy. The ranking might change. Not surprisingly, in almost all cases the VARMA algorithms do better than the benchmark long VAR. In most cases, the VARMA algorithms also display smaller MSE ratios than a VAR chosen by AIC. 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 modelling. Increasing the forecast horizon, does reduce the differences between the different algorithms. The same is true when more observations are available. Increasing the complexity in terms of Kronecker indices does have minor effects. The forecasts obtained by the CCA method are often comparable but often also inferior to the forecasts obtained by other algorithms. In particular, the CCA forecasts are often inferior for the one-step forecast horizons. The 2SLS estimator yields usually better forecasts than the CCA forecasts and comparable but sometimes slightly worse forecast than the other VARMA algorithms. The GLS and the IOLS procedure do quite well in forecasting depending on the specific DGP and number of observations. The 3SLS procedure, however, seems to be slightly preferable. The MLE method is always superior to all simple algorithms apart from one case, DGP III, LNMAEV with $T = 100$, where its MSE is roughly three times as large as the MSE of the other VARMA algorithms. In this case, the MSE ratio for the MLE procedure is not shown on the graph since this would imply loosing important details in other parts. In general, however, the differences are small, in particular in comparison to

the rather large differences in terms of parameter estimation accuracy. In sum, the ranking of the different algorithms becomes less clear when forecasting is the objective. While the VARMA methods do generally better than the VARs and the CCA method, the differences are often small. For the simulated processes, 3SLS 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 VARMA modelling. When the impulse response horizon is increased, VARMA modelling 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 example, for the rather simple DGP I, there are little advantages of VARMA modelling apart from the LPMAEV and LNMAEV parameterizations. For the other DGPs there are in principle considerable advantages provided that the right algorithm is chosen and the process is correctly specified. Furthermore, the VARMA algorithms differ much more at horizon $h = 1$. Increasing the sample size has no important effect on the ranking of the algorithms. First, the CCA method seems to be inferior to the VARMA algorithms for all DGPs and both horizons. Occasionally, CCA is worse than the VAR chosen by AIC. The 2SLS algorithm estimates the impulse responses with comparable or slightly worse accuracy than the other VARMA algorithms. Only for DGP II the impulse response estimates obtained by 2SLS are as precise as the estimates obtained by other algorithms. Also the results for the impulse response estimates obtained by GLS are mixed. In some cases, such as DGP I with large moving-average eigenvalues, GLS is performing quite well but in most other cases GLS is inferior to IOLS or 3SLS. In fact, these two algorithms estimate the impulse response function best in most of the cases. While the performance of IOLS in this respect depends still on the specific DGP, 3SLS is almost always the preferable method. Furthermore, even though IOLS is often the second-best method, the difference to 3SLS can be considerable, in particular for higher-order processes. In sum, the 3SLS procedure is by far preferable, independent of the specific DGP at hand. Generally, the impulse response estimates obtained by MLE are much more precise than the corresponding

estimates obtained by the 3SLS algorithm. These results correspond to the statements made above about the algorithms' relation in terms of parameter estimation accuracy. Overall, VARMA modelling 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.

In sum, VARMA modelling can be advantageous. 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. Apart from forecasting, there are large differences between the algorithms. Overall, the algorithm, which is closest to maximum likelihood estimation, 3SLS, seems to be superior to any other of the simpler estimation algorithms. In particular, when the complexity of the simulated systems increases, 3SLS is the only algorithm that almost always outperforms the benchmark VARs in terms of accuracy of the estimated impulse responses. A concern, however is the instability of the algorithm in the presence of large eigenvalues of the moving-average polynomial. Even though full-information maximum likelihood would be the ideal algorithm, 3SLS is performing quite well in comparison not only to the alternative simple VARMA algorithms but also in comparison to the benchmark VARs. However, as the algorithm is implemented here, it is still not stable enough in order to be used in a automatic fashion because of the non-invertibility problem. Given the simplicity of the used DGPs and that complications such as specification, outliers etc. are neglected, these results suggest that the 3SLS algorithm would have to be improved considerably in order to create an algorithm that returns accurate estimates in almost all cases.

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 modelling.

It has been shown in the simulations that there are situations where the investigated algorithms do not perform very well. There is a rough trade-off between the technical reliability of the algorithms and the quality of the estimates. With respect to the accuracy of the parameter estimates, the iterative least squares procedure of Kapetanios (2003) and the simple least squares procedure of Hannan & Kavalieris (1984*a*) seem to perform relatively well for smaller processes with small eigenvalues of the moving-average part. However, they can be quite imprecise relative to exact maximum likelihood for higher dimensional processes and in particular for processes with large eigenvalues in the moving-average part.

If the purpose of time series analysis is forecasting, the methods perform approximately comparable though few can reach or outperform the forecasting power of exact maximum likelihood. The gains from using VARMA models in contrast to VARs appear to be relatively small. Also, in this case the procedure of Hannan & Kavalieris (1984*a*) turned out to be preferable over the other simpler estimation algorithms.

The true impulse responses are estimated poorly by most algorithms given the benchmark of a long VAR. Again, the procedure of Hannan & Kavalieris (1984*a*) is potentially quite advantageous. Also the iterative least squares procedure of Kapetanios (2003) is performing well in this respect. Nevertheless, the algorithms cannot reach the precision of the exact maximum likelihood procedure.

It turns out, that the only simple procedure that reliably gave significantly better results than the benchmark VARs in terms of the accuracy of the derived forecasts and impulse response estimates, is the procedure which is closest to maximum likelihood, namely the procedure of Hannan & Kavalieris (1984*a*). However, this procedure is also the most unreliable procedure in technical terms, in that it often yields estimated models which are not invertible. Given the simplicity of the simulated data generating processes, the algorithm would have to be improved considerably in order to make it a standard tool for applied researchers.

A reliable and accurate algorithm for the estimation of VARMA models still remains to be developed. This study suggests that there are potentially considerable gains from VARMA modelling. 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.

Aoki, M. (1989), *State Space Modeling of Time Series*, Springer-Verlag, Berlin.

Bauer, D. (2005*a*), 'Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs.', *Journal of Time Series Anaysis* **26**(5), 631–668.

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

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., Peternell, K. & Scherrer, W. (1995), 'Consistency and relative efficiency of subspace methods', *Automatica* **31**, 1865–1875.

Desai, U. B., Pal, D. & Kirkpatrick, R. D. (1985), 'A realization approach to stochastic model reduction', *International Journal of Control* **42**(4), 821–838.

Dufour, J.-M. & Pelletier, D. (2004), 'Linear estimation of weak VARMA models with a macroeconomic application'. Université de Montréal and North Carolina State University, Working Paper.

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. & Sargent, T. J. (2005), 'A,B,C's (and D)'s for understanding VARs'. NBER Technical Working Paper 308, May draft.

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.

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.

Koreisha, S. & Pukkila, T. (1990a), 'Linear methods for estimating ARMA and regression models with serial correlation.', *Communications in Statistics-Simulation* **19**, 71–102.

Larimore, W. E. (1983), System Identification, Reduced-Order Filters and Modeling via Canonical Variate Analysis, *in* H. S. Rao & P. Dorato, eds, 'Proc. 1983 Amer. Control Conference 2'.

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

Lütkepohl, H. & Poskitt, D. S. (1996), 'Specification of Echelon-Form VARMA Models', *Journal of Business & Economic Statistics* **14**(1), 69–79.

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

Newbold, P. & Granger, C. W. J. (1974), 'Experiences with forecasting univariate time series and combination of forecasts', *Journal of the Royal Statistical Society* **A137**, 131–146.

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.

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

A Equivalence between VARMA and State Space Representations

This discussion serves as an illustration and is based on the corresponding sections in Aoki's (1989) book.[7] It is not claimed, for example, that the following state space representation of a VARMA model is especially meaningful. The point is simply to demonstrate that a VARMA model *can* be written in state space form. Suppose that a multiple time series $y_t = (y_{1t}, \ldots, y_{Kt})'$ of dimension K satisfies a VARMA(p, q) model given by

$$y_t = \sum_{i=1}^{p} A_i y_{t-i} + u_t + \sum_{i=1}^{q} M_i u_{t-i},$$

where $A_0 = I_K$ is assumed for simplicity. This process can be written as a state space model by defining

$$A := \left[\begin{array}{cccc|cccc} A_1 & \ldots & \ldots & A_p & M_1 & M_2 & \ldots & M_q \\ I_K & 0 & & & 0 & 0 & & \\ & \ddots & \ddots & & 0 & \ddots & & \\ & & I_K & 0 & & & & \\ \hline 0 & 0 & & & 0 & & & \\ 0 & \ddots & & & I_K & 0 & & \\ & & & & & \ddots & \ddots & \\ & & & 0 & & & I_K & 0 \end{array}\right], ((p+q)K \times (p+q)K),$$

$$B' := \left[\begin{array}{ccccccc} I_K : & 0 : & \ldots & I_K : & \ldots : 0 \end{array}\right], (K(p+q) \times K),$$

$$C := \left[\begin{array}{ccccccc} A_1 : & \ldots & : A_p & : M_1 : & \ldots & : M_q \end{array}\right], (K \times K(p+q)).$$

[7] See also the book of Hannan & Deistler (1988) for an extensive discussion on the relation between state space and VARMA models.

The state space model is of the form

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \\ y_t &= Cx_t + u_t, \end{aligned}$$

with a state vector given by

$$x_t = \begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-p} \\ u_{t-1} \\ \vdots \\ u_{t-q} \end{bmatrix}, ((p+q)K \times 1).$$

Given a state space model of order n for a K-dimensional process, let the characteristic polynomial of the system matrix A be $|A - \lambda I_n| = c_0\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \ldots + c_n,\ c_0 = 1$. Multiply the observation equation for $t, \ldots, t+n$ with the coefficients $c_i,\ i = 0, \ldots, n$, in the following way

$$\begin{aligned} c_n y_t &= c_n(Cx_t + u_t), \\ c_{n-1} y_{t+1} &= c_{n-1}(CAx_t + CBu_t + u_{t+1}), \\ &\vdots \\ y_{t+n} &= CA^n x_t + CA^{n-1}Bu_t + \ldots + CBu_{t+n-1} + u_{t+n}, \end{aligned}$$

where the right hand side has been obtained by recursive substitution. Summing up these equations one obtains

$$y_{t+n} + c_1 y_{t+n-1} + \ldots + c_n y_t = C(A^n + c_1 A^{n-1} + \ldots + c_n I_n)x_t + \sum_{i=0}^{n} D_i u_{t+i}$$

where $D_i = c_{n-i} I_K + \sum_{k=1}^{n-i} c_{n-i-k} C A^{k-1} B$. According to the Cayley - Hamilton theorem, the matrix polynomial in A vanishes, $A^n + c_1 A^{n-1} + \ldots + c_n I_n = 0$ (Aoki 1989). One obtains therefore the following VARMA representation

$$y_{t+n} + c_1 y_{t+n-1} + \ldots + c_n y_t \quad = \quad \sum_{i=0}^{n} D_i u_{t+i}.$$

B Figures and Tables

Table 1: Parameter Values

DGP		Parameters	λ^{ar}	λ^{ma}
DGP I	MEV	$\alpha_1 = 0.2,\ m_{11,1} = 0.25$ $m_{22,1} = -0.1$	0.2	0.1, 0.05
	LPAREV	$\alpha_1 = 0.9,\ m_{11,1} = 0.25$ $m_{22,1} = -0.1$	0.9	0.1, 0.05
	LNAREV	$\alpha_1 = -0.9,\ m_{11,1} = 0.25$ $m_{22,1} = -0.1$	-0.9	0.1, 0.05
	LPMAEV	$\alpha_1 = 0.2,\ m_{11,1} = 0.98$ $m_{22,1} = 0.52$	0.2	0.9, 0.6
	LNMAEV	$\alpha_1 = 0.2,\ m_{11,1} = -0.52$ $m_{22,1} = -0.98$	0.2	-0.9, -0.6
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
	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
	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
	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 0.13
	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 0.13
DGP III	MEV	$\alpha_{11,1} = 0.7,\ m_{22,1} = -0.6$	0.7	-0.6, 0.5
	LPAREV	$\alpha_{11,1} = 0.9,\ m_{22,1} = -0.6$	0.9	-0.6, 0.5
	LNAREV	$\alpha_{11,1} = -0.9,\ m_{22,1} = -0.6$	-0.9	-0.6, 0.5
	LPMAEV	$\alpha_{11,1} = 0.7,\ m_{22,1} = 0.9$	0.7	0.9, 0.5
	LNMAEV	$\alpha_{11,1} = 0.7,\ m_{22,1} = -0.9$	0.7	-0.9, 0.5
DGP IV	MEV	$\alpha_{11,1} = 0.5,\ m_{55,1} = -0.6$	0.5, 0 ± i 0.57	-0.6, -0.4 ± i 0.67
	LPAREV	$\alpha_{11,1} = 0.9,\ m_{55,1} = -0.6$	0.9, 0 ± i 0.57	-0.6, -0.4 ± i 0.67
	LNAREV	$\alpha_{11,1} = -0.9,\ m_{55,1} = -0.6$	-0.9, 0 ± i 0.57	-0.6, -0.4 ± i 0.67
	LPMAEV	$\alpha_{11,1} = 0.5,\ m_{55,1} = 0.9$	0.5, 0 ± i 0.57	0.9, -0.4 ± i 0.67
	LNMAEV	$\alpha_{11,1} = 0.5,\ m_{55,1} = -0.9$	0.5, 0 ± i 0.57	-0.9, -0.4 ± i 0.67

Varied parameter values and corresponding eigenvalues of the autoregressive and the moving-average parts for the different data generating processes.

Table 2: Non-invertible Estimated Models, $T = 100$

DGP		2SLS	3SLS	GLS	IOLS	CCA
DGP I	MEV	0.0	0.0	0.0	0.0	0.0
	LPAREV	0.0	0.0	0.0	0.0	0.0
	LNAREV	0.0	0.0	0.0	0.0	0.1
	LPMAEV	1.7	4.9	0.0	0.5	0.6
	LNMAEV	0.8	8.9	0.0	0.5	0.2
DGP II	MEV	0.2	3.3	0.0	0.7	0.1
	LPAREV	0.2	1.1	0.0	0.4	0.1
	LNAREV	0.0	1.0	0.0	0.1	0.4
	LPMAEV	1.0	4.1	0.0	2.5	0.1
	LNMAEV	0.7	8.9	0.0	3.5	0.0
DGP III	MEV	0.2	3.9	0.3	0.3	0.0
	LPAREV	0.2	3.6	0.1	0.2	0.1
	LNAREV	0.2	2.5	0.2	0.1	0.0
	LPMAEV	2.8	6.2	0.3	1.0	0.5
	LNMAEV	1.5	11.3	0.2	0.5	0.1
DGP IV	MEV	0.0	3.4	0.1	0.0	0.0
	LPAREV	0.1	1.6	0.1	0.0	0.1
	LNAREV	0.2	1.5	0.1	0.0	0.1
	LPMAEV	1.1	9.2	0.3	0.6	0.3
	LNMAEV	1.7	10.0	0.3	0.2	0.2

Frequency of cases in percentage when the algorithms returned non-invertible models or, in case of the CCA algorithm, yielded models that violated the minimum phase assumption.

Table 3: Non-invertible Estimated Models, $T = 200$

DGP		2SLS	3SLS	GLS	IOLS	CCA
DGP I	MEV	0.0	0.0	0.0	0.0	0.0
	LPAREV	0.0	0.0	0.0	0.0	0.0
	LNAREV	0.0	0.0	0.0	0.0	0.0
	LPMAEV	0.8	1.6	0.0	0.2	0.0
	LNMAEV	0.1	5.0	0.0	0.0	0.1
DGP II	MEV	0.0	0.1	0.0	0.1	0.0
	LPAREV	0.0	0.1	0.0	0.0	0.0
	LNAREV	0.0	0.2	0.0	0.0	0.0
	LPMAEV	0.0	0.2	0.0	0.3	0.0
	LNMAEV	0.1	7.3	0.0	1.0	0.0
DGP III	MEV	0.0	0.5	0.2	0.0	0.0
	LPAREV	0.0	0.4	0.2	0.0	0.0
	LNAREV	0.0	0.1	0.0	0.0	0.0
	LPMAEV	0.7	1.8	0.4	0.5	0.0
	LNMAEV	0.3	4.8	0.0	0.5	0.0
DGP IV	MEV	0.0	0.1	0.1	0.0	0.0
	LPAREV	0.0	0.0	0.3	0.0	0.0
	LNAREV	0.0	0.5	0.0	0.0	0.0
	LPMAEV	0.2	3.6	0.1	0.0	0.0
	LNMAEV	0.2	3.5	0.2	0.0	0.2

Frequency of cases in percentage when the algorithms returned non-invertible models or, in case of the CCA algorithm, yielded models that violated the minimum phase assumption.

Table 4: Total Estimation Failures, $T = 100$

DGP		2SLS	3SLS	GLS	IOLS	CCA
DGP I	MEV	0.0	0.0	0.0	0.0	0.0
	LPAREV	0.0	0.0	0.0	0.0	0.0
	LNAREV	0.0	0.0	0.0	0.0	0.1
	LPMAEV	1.7	4.9	0.0	5.0	0.6
	LNMAEV	0.8	8.9	0.0	3.6	0.2
DGP II	MEV	0.2	3.3	0.0	0.7	0.1
	LPAREV	0.2	1.1	0.0	0.4	0.1
	LNAREV	0.0	1.0	0.0	0.1	0.4
	LPMAEV	1.0	4.1	0.0	2.5	0.1
	LNMAEV	0.7	8.9	0.0	3.5	0.0
DGP III	MEV	0.2	3.9	0.3	0.9	0.0
	LPAREV	0.2	3.6	0.1	0.6	0.1
	LNAREV	0.2	2.5	0.2	0.7	0.0
	LPMAEV	2.9	6.2	0.3	3.6	0.5
	LNMAEV	1.5	11.3	0.2	1.6	0.1
DGP IV	MEV	0.0	3.4	0.1	2.3	0.0
	LPAREV	0.1	1.6	0.1	0.7	0.1
	LNAREV	0.2	1.5	0.1	0.4	0.1
	LPMAEV	1.1	9.2	0.3	4.7	0.3
	LNMAEV	1.7	10.0	0.3	3.3	0.2

Frequency of cases in percentage when the algorithms returned non-invertible models, did not converge, or returned an extreme outlier.

Table 5: Total Estimation Failures, $T = 200$

DGP		2SLS	3SLS	GLS	IOLS	CCA
DGP I	MEV	0.0	0.0	0.0	0.0	0.0
	LPAREV	0.0	0.0	0.0	0.0	0.0
	LNAREV	0.0	0.0	0.0	0.0	0.0
	LPMAEV	0.8	1.6	0.0	1.1	0.0
	LNMAEV	0.1	5.0	0.0	0.9	0.1
DGP II	MEV	0.0	0.1	0.0	0.1	0.0
	LPAREV	0.0	0.1	0.0	0.0	0.0
	LNAREV	0.0	0.2	0.0	0.0	0.0
	LPMAEV	0.0	0.2	0.0	0.3	0.0
	LNMAEV	0.1	7.3	0.0	1.0	0.0
DGP III	MEV	0.0	0.5	0.2	0.0	0.0
	LPAREV	0.0	0.4	0.2	0.1	0.0
	LNAREV	0.0	0.1	0.0	0.0	0.0
	LPMAEV	0.7	1.8	0.4	0.9	0.0
	LNMAEV	0.3	4.8	0.8	0.6	0.0
DGP IV	MEV	0.0	0.1	0.1	0.1	0.0
	LPAREV	0.0	0.0	0.3	0.0	0.0
	LNAREV	0.0	0.5	0.0	0.0	0.0
	LPMAEV	0.2	3.6	0.1	0.2	0.0
	LNMAEV	0.2	3.5	0.2	0.1	0.2

Frequency of cases in percentage when the algorithms returned non-invertible models, did not converge, or returned an extreme outlier.

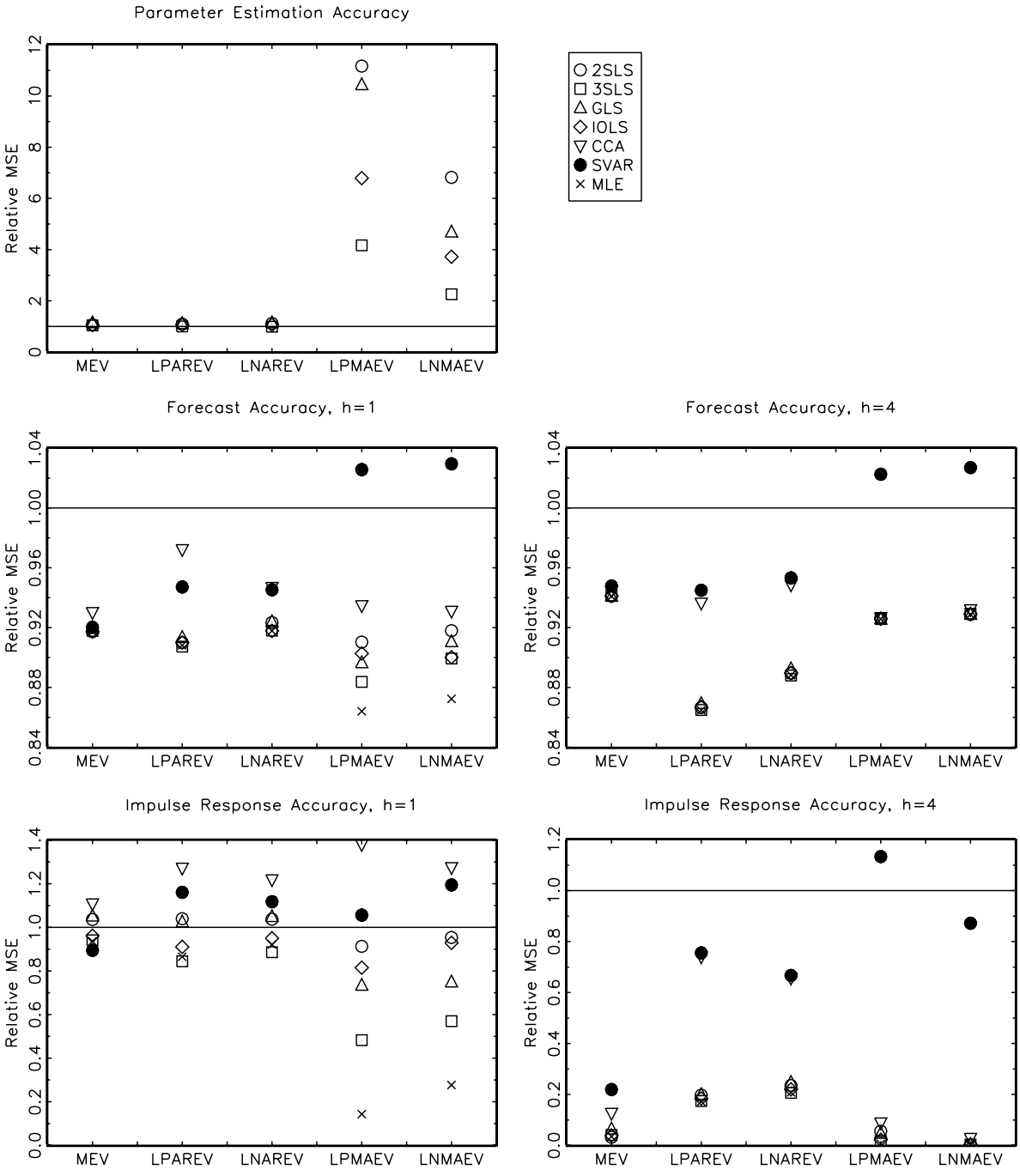

Figure 1: MSE ratios for DGP I with $T = 100$.

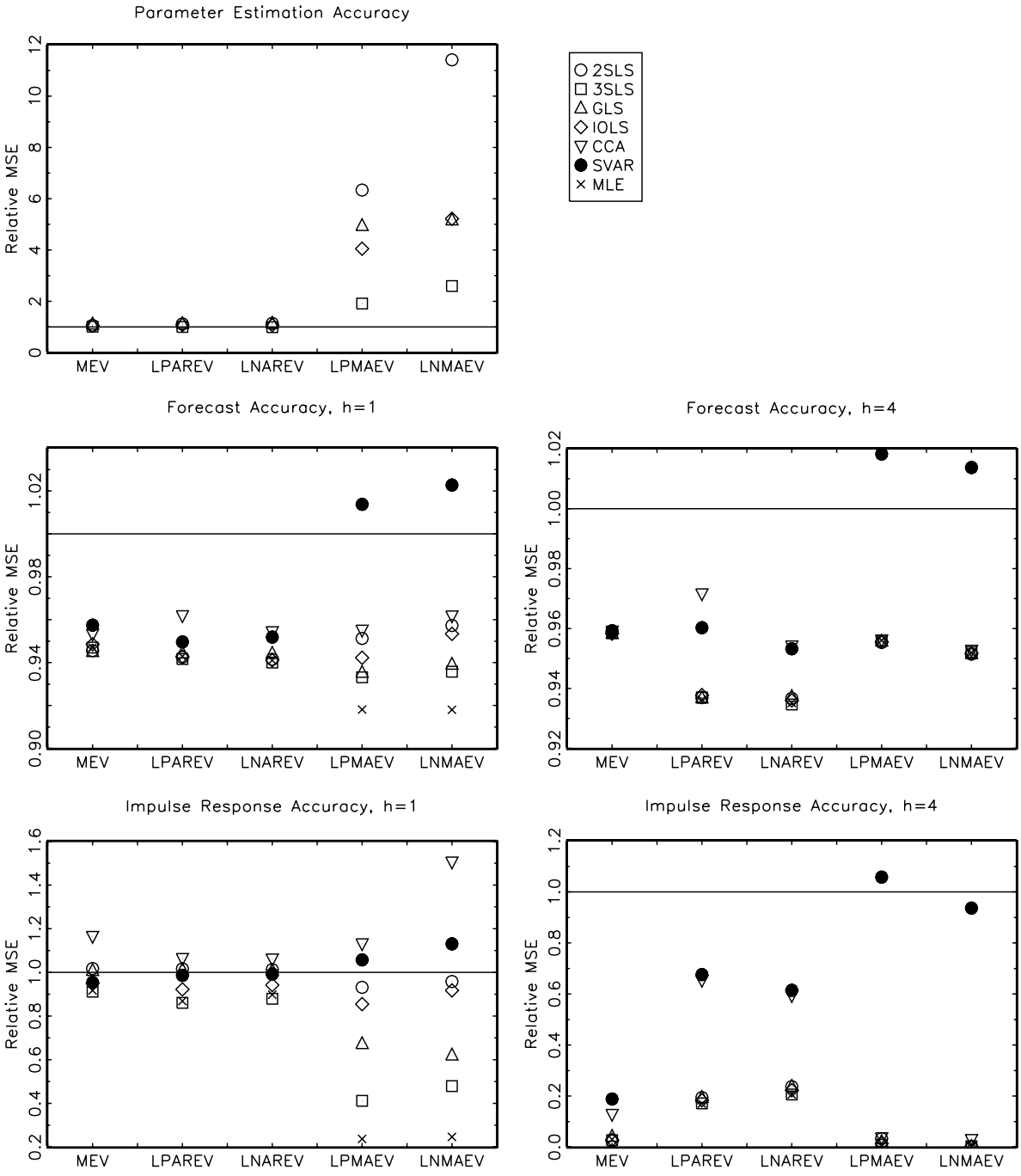

Figure 2: MSE ratios for DGP I with $T = 200$.

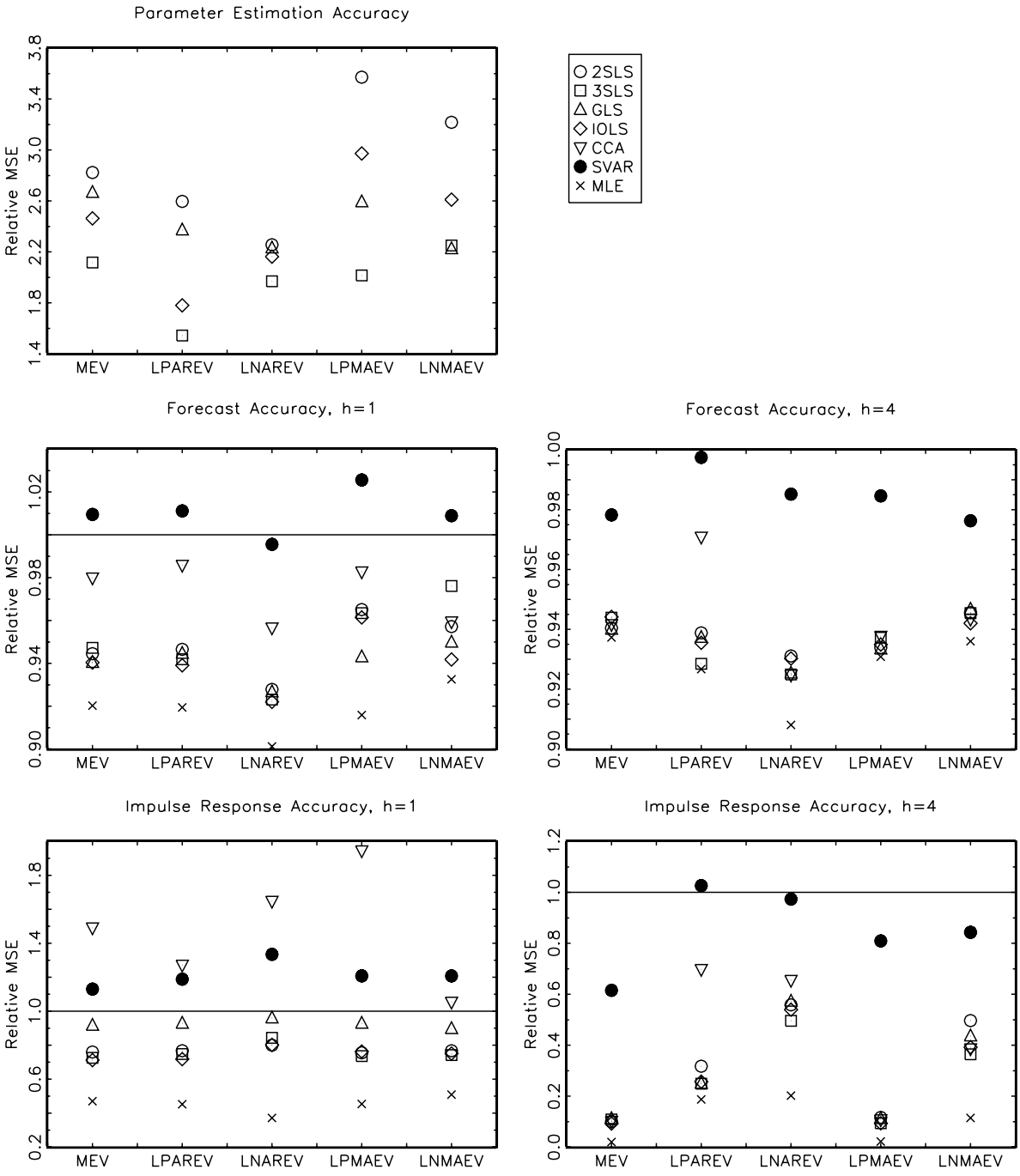

Figure 3: MSE ratios for DGP II with $T = 100$.

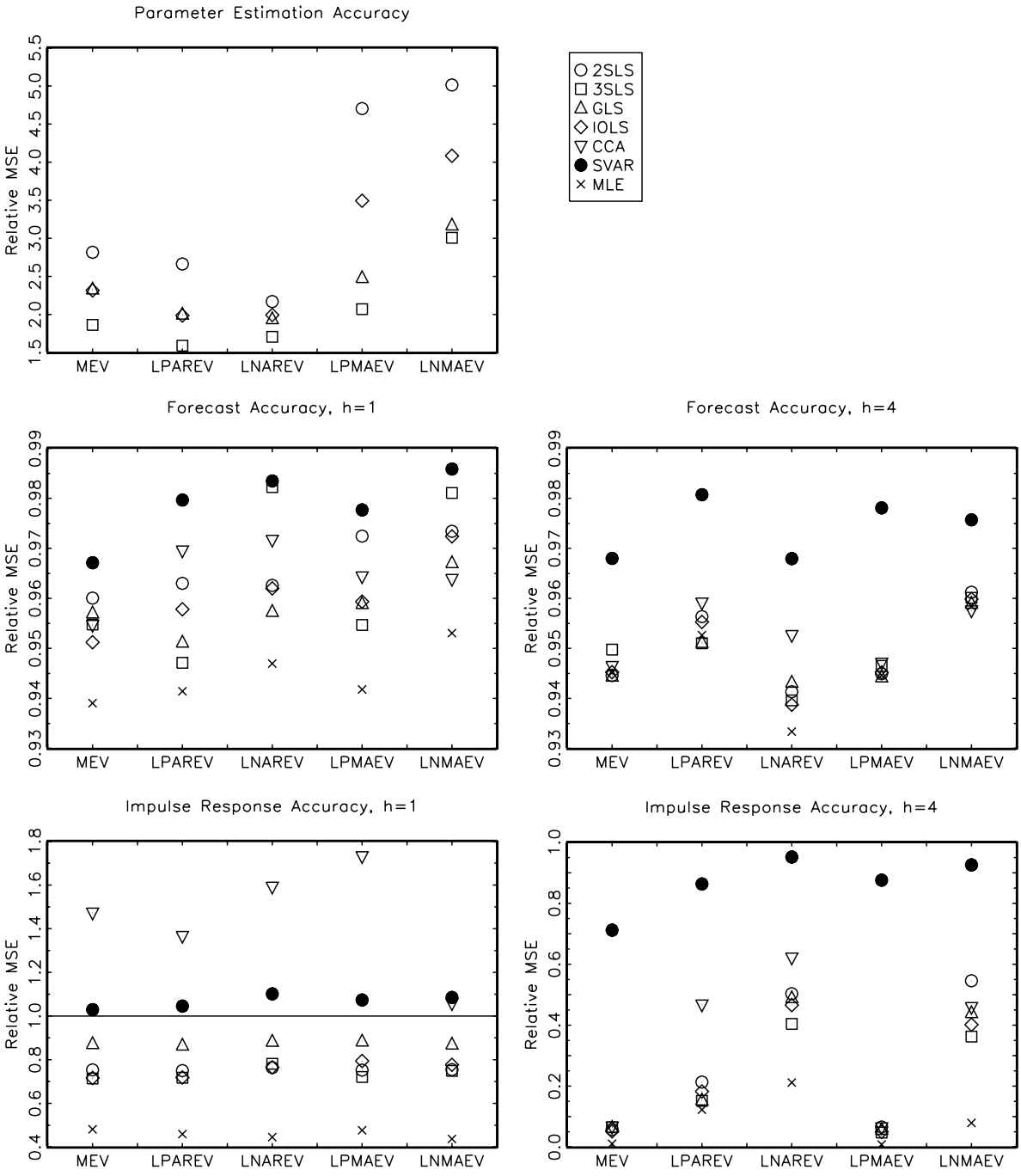

Figure 4: MSE ratios for DGP II with $T = 200$.

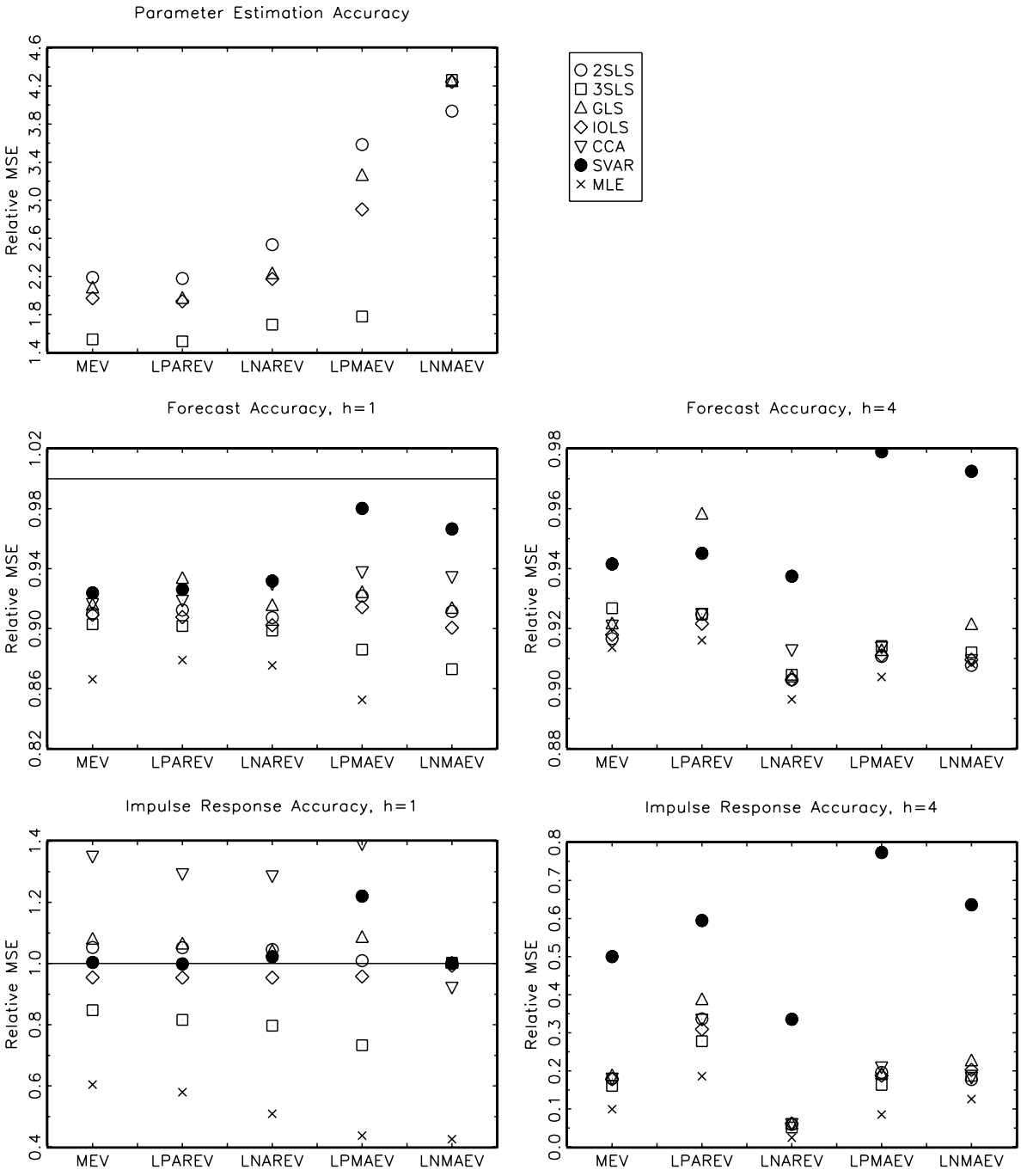

Figure 5: MSE ratios for DGP III with $T = 100$.

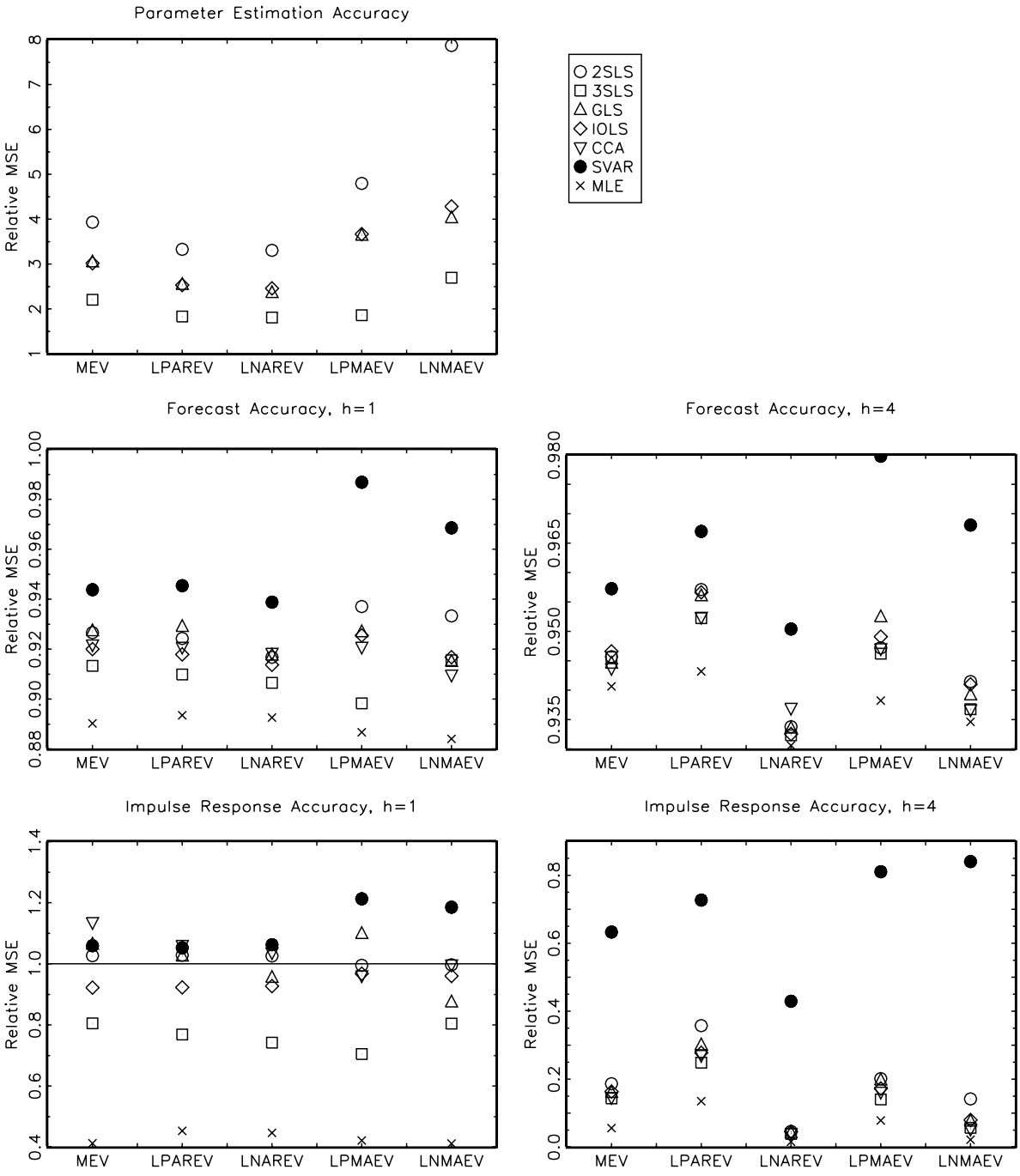

Figure 6: MSE ratios for DGP III with $T = 200$.

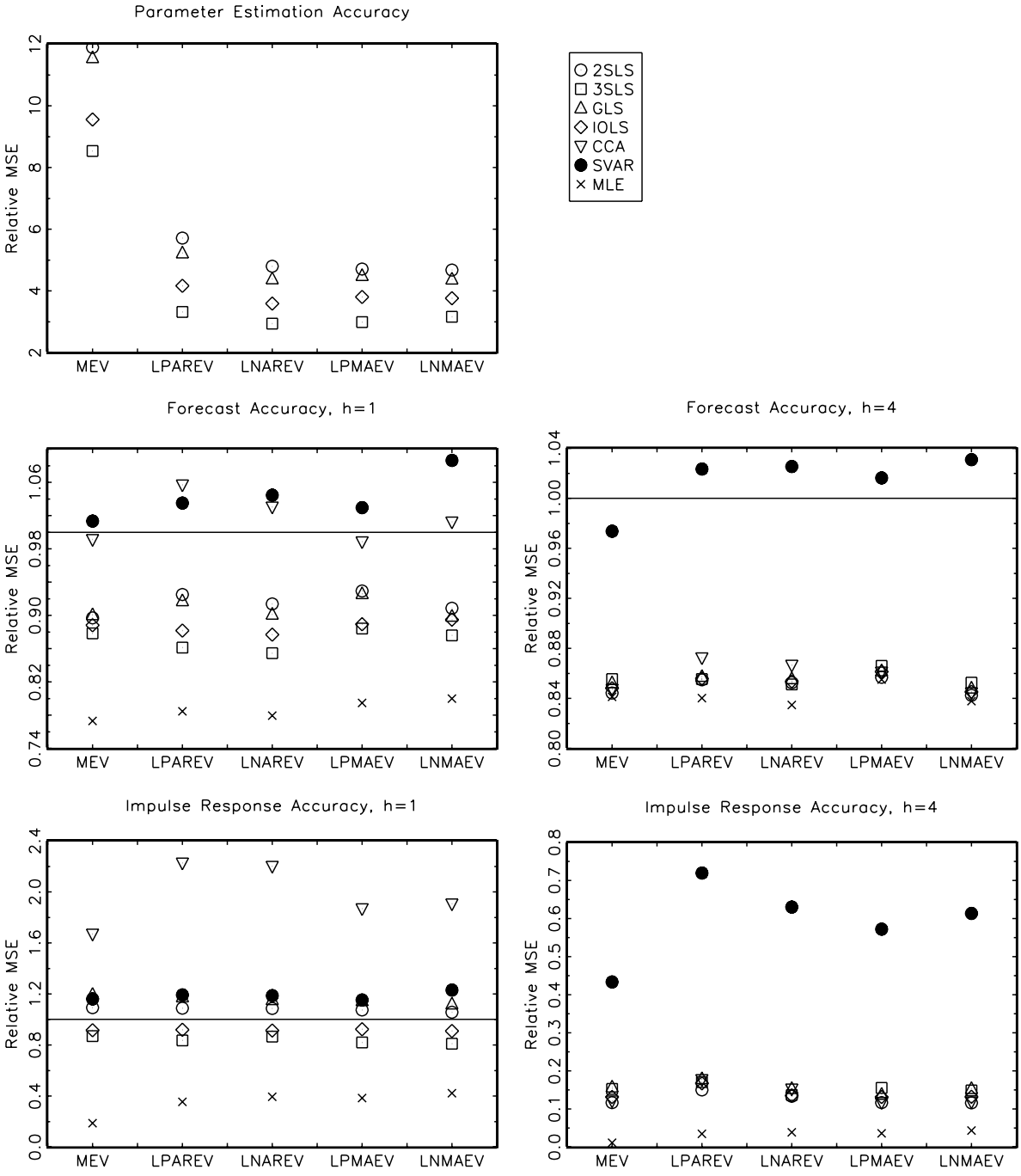

Figure 7: MSE ratios for DGP IV with $T = 100$.

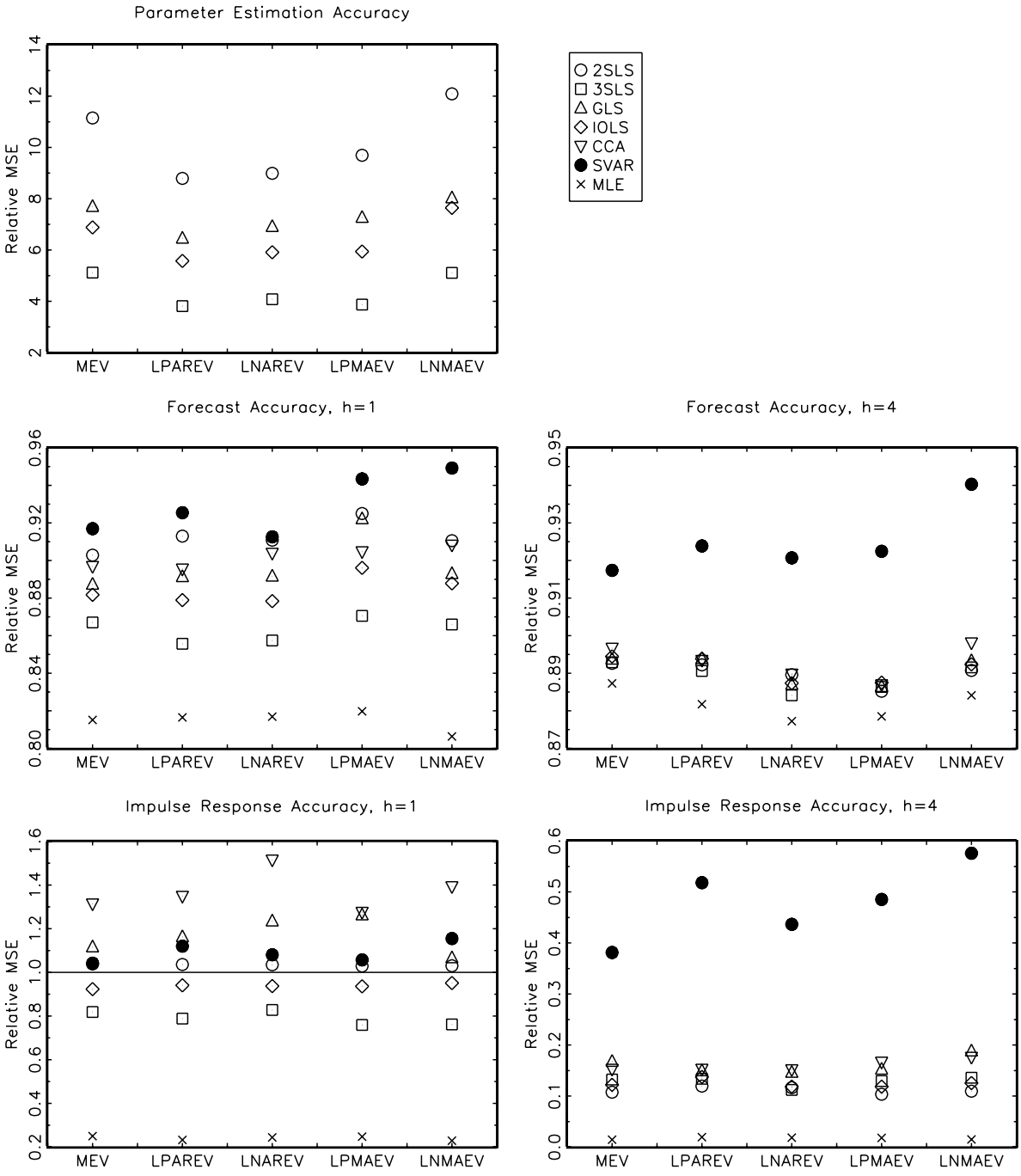

Figure 8: MSE ratios for DGP IV with $T = 200$.