European
University
Institute

Department of Economics

Three Essays in Time Series Econometrics

Christian Kascha

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Florence, September 2007

EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

Three Essays in Time Series Econometrics

Christian Kascha

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Examining Board:
Prof. Helmut Lütkepohl; EUI, Supervisor
Prof. Anindya Banerjee, EUI
Prof. Helmut Herwartz, University of Kiel
Prof. Pentti Saikkonen, University of Helsinki

To my parents and to Irene

Acknowledgements

This thesis would not have been possible without the support, comments and criticism of many people. In particular, I would like to thank Professor Helmut Lütkepohl for his constant support and supervision. It was a pleasure to have him as a supervisor. I am grateful to my second reader, Professor Anindya Banerjee, and to the members of my thesis committee, Professor Helmut Herwartz and Professor Pentti Saikkonen, for their advice and comments. I would like to thank Professor Morten O. Ravn for many helpful comments and discussions. It was a pleasure to work with Karel Mertens on a joint paper which is chapter two of this thesis.

I also would like to thank my parents and Irene for their constant support and for bearing a considerable part of the social costs of this thesis. I am grateful to many friends and colleagues. I would like to thank Judith Ay, Andrea Barone, Zeno Enders, Lapo Filistrucchi, Andrea Herrmann, Matthias Hertweck, Dejan Krusec, Christopher Milde, Markus Poschke, Elvira Prades, Katrin Rabitsch, Clara Solanes, Claudia Trentini and Sanne Zwart.

All errors, of course, are mine.

Contents

Part I

Introduction

Introduction

This thesis deals with different topics in time series econometrics that belong, broadly speaking, to the area of macroeconometrics. That is, topics and methods are investigated which are of interest to applied researchers that want to analyze the behavior of aggregate measurements of the economy by means of time series data. For instance, macroeconomic time series often display specific nonlinear characteristics. Chapter 1 applies one of the recently proposed techniques to capture nonlinearity to U.S. interest rate data. Another important area of macroeconomic research is the identification of structural shocks that have an economically meaningful interpretation, in contrast to prediction errors. Chapter 2 compares some alternative structural estimators in the context of a recent discussion on the reliability of standard structural estimators. Finally, the third chapter compares different estimators of so-called vector autoregressive moving-average (VARMA) models that are a potential alternative to vector autoregressive models which are used predominantly in applied macroeconomic research. In the following, I will briefly describe each chapter in more detail.

Chapter 1 reviews some simple extensions of the threshold cointegration models forwarded by Balke and Fomby (1997). These are models designed for non-stationary series with nonlinear dynamics that are tight together by a long-run equilibrium relationship. I apply one of these models to the U.S. interest rate spread and show that there is evidence in favor of a threshold model of this kind. The model suggests that the spread is probably not mean-reverting in a large band around the equilibrium. However, the model is not found to provide better forecasts relative to a linear benchmark vector error-correction model.

In chapter 2 Karel Mertens and I compare different structural estimators in the context of a recent discussion on the reliability of long-run identified structural vector autoregressions (SVARs). Several authors have suggested that SVARs fail partly because they are finite-order approximations to infinite-order processes. In the second chapter we estimate VARMA and state space models, which are not misspecified, using simulated data and compare true with estimated impulse responses of hours worked to a technology shock. We find little gain from using VARMA models. However, state space models do outperform SVARs. In particular, subspace methods consistently yield lower mean squared errors, although even these estimates remain too imprecise for reliable inference. Our findings indicate that long-run identified SVARs perform weakly not because of the finite-order approximation. Instead, we find that the simulated processes used in the previous studies are nearly violating the most basic assumptions on which long-run identification schemes are based.

Chapter 3 compares different estimation methods for VARMA models. Classical Gaussian maximum likelihood estimation 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 (VAR) models in macroeconomic research. Therefore, several other, simpler estimation methods have been proposed in the literature. In chapter 3 these methods are compared by means of a Monte Carlo study. Different evaluation criteria are used to judge the relative performances of the algorithms. The obtained results suggest that the algorithm of Hannan and Kavalieris (1984a) is the only algorithm that reliably outperforms the other algorithms and the benchmark VARs. However, the procedure is technically not very reliable and therefore would have to be improved in order to make it an alternative tool for applied researchers.

Part II

Chapters

Chapter 1

Threshold Cointegration and the Interest Rate Spread

1.1 Introduction

It is widely believed that many economic variables behave and are related in a nonlinear fashion. So far econometric theory has been primarily concerned with linear models that are unable to generate some apparent, nonlinear features of the data such as asymmetry, limit cycles or amplitude dependent frequency. Nonlinear models have only recently received much attention in the time series literature. Important examples in the econometric literature are Markov-Switching models (Goldfeld and Quandt, 1973; Hamilton, 1989), threshold autoregressive models (Tong, 1978; Chan and Tong, 1986), smooth transition models (Bacon and Watts, 1971; Teräsvirta and Anderson, 1992) and artificial neural networks (see, e.g., Kuan and White, 1994).

In particular, threshold models have been first proposed by Tong (1978) as a simple but potentially powerful alternative to linear econometric models. In threshold autoregressive (TAR) models the relation between the dependent variable and its lags depends on whether another variable, a so-called threshold variable, falls in a certain interval or *regime*. The limits of these intervals are called thresholds. Conditional on a certain value of the threshold variable, the relation between the dependent variable and its lags is linear. Therefore, the linear autoregressive model is a special case of a threshold autoregressive model with only one regime. That is, the threshold model nests the standard autoregressive model. The property that the threshold model is linear conditional on the threshold variable makes is analytically more tractable than comparable models that do not share this property, in general. Another advantage is that testing linearity is straightforward. Furthermore, even though the model is not overly complicated, it is quite flexible in that it allows to approximate different forms of nonlinearity. The asymptotic theory for the estimators of threshold models has long lagged behind until recently (see, e.g., Hansen, 1999, 2000). Still today the properties of these models such as stationarity, existence of moment etc. are not fully understood.

The extension to multivariate models is straightforward. For an account on threshold vector autoregressive models see the paper of Tsay (1998). Threshold cointegration models have been proposed by Balke and Fomby (1997) to model jointly non-stationarity and nonlinearity. These models have been found useful for modelling monetary policy variables (Baum and Karasulu, 1998), testing the purchasing power parity (Obstfeld and Taylor, 1997; Taylor, 2001; Lo and Zivot, 2001) and modelling nonlinear mean reversion in the term structure (Seo, 2003). Recently, Hansen and Seo (2002) proposed a method to implement maximum-likelihood estimation of the threshold cointegration model in the bivariate case and developed a test for the presence of a threshold effect. However, there are several open questions for future research related to the maximum likelihood estimator of threshold cointegrated systems. For instance, one would like to allow for models with multiple cointegrating vectors and the asymptotic distribution theory for all the estimated parameters, including the threshold estimates, has yet to be developed. Another fundamental question is how to restrict the parameters of a threshold cointegration model such that certain deterministic properties of the data such as trends or seasonality are captured properly in these models.

This paper is a brief review of some simple generalizations of the threshold cointegration models proposed by Balke and Fomby (1997) and I discuss how one rules out deterministic trends in these very special cases. I then apply one of these models to the U.S. term structure and review the relevant tools for the specification and estimation of these models while analyzing the series. The application investigates the validity of the expectation hypothesis which is interesting to practitioners and economists as well in that the implied cointegration relation might provide better forecasts and gives insight in the efficiency of financial markets. I find that there is evidence for threshold nonlinearity though the proposed model does not yield superior forecasts relative to the linear VECM.

The remaining part of the paper is organized as follows. First, I review the particular threshold models and discuss how these models can be interpreted and estimated. Second, I apply one of them to a bivariate series of U.S. interest rates, provide evidence for threshold nonlinearity and discuss the essential specification tools and, third, I comment on the estimation results as well as possible theoretical extensions and conclude.

1.2 Econometric Methods

In this section I review some generalizations of the models proposed by Balke and Fomby (1997) and discuss how a linear trend is ruled out in these models. Careful handling of the deterministic terms is important in modelling non-stationary time series. Estimation methods that impose restrictions on the intercepts and trends such that the actual model replicates main features of the data is likely to improve the fit of the model and to increase estimation efficiency. In threshold models, however, it is far from obvious how to deal with these issues. For illustration, a simple, univariate threshold model with no intercept like

$$y_t = \begin{cases} 0.9y_{t-1} + \varepsilon_t, & y_{t-1} > 0 \\ 0.2y_{t-1} + \varepsilon_t, & y_{t-1} \leq 0 \end{cases}$$

with $\varepsilon_t \sim iid\, N(0, \sigma^2)$ generates a series with nonzero mean. The reason is that though the shocks are symmetrically distributed around zero they are much more persistent in one direction than in the other. This simple process exemplifies the difficulties that are encountered when modelling nonlinear integrated series and that, in particular, restrictions on the intercept in threshold cointegrated systems are not easily implemented. Therefore, I consider only threshold cointegrated systems in which it is very easy to restrict the intercept term properly. Precisely, I consider the following three-regime model for a bivariate series $y_t = (y_{1t},\ y_{2t})'$ with p lags in differences and one cointegrating relation given by

$$y_t \quad = \quad \mu + x_t,$$

where μ denotes a constant (2×1) vector and x_t is a non-stationary process that will be defined below. I require for my purpose that $E(\Delta x_t) = 0$ such that $E(\Delta y_t) = 0$ and thus the observed series does not display a deterministic downward or upward movement. This is achieved by considering a symmetric but nonlinear model in first differences. Precisely, the stochastic part, x_t, has a threshold vector error correction (TVECM) representation as

$$\Delta x_t = \begin{cases} \alpha^{(2)}\beta' x_{t-1} + \Gamma^{(2)}\Delta X_{t-1} + u_t, & |\beta' x_{t-1}| > \gamma \\ \alpha^{(1)}\beta' x_{t-1} + \Gamma^{(1)}\Delta X_{t-1} + u_t, & |\beta' x_{t-1}| \leq \gamma, \end{cases}$$

where $\alpha^{(j)}$, $j = 1, 2$, denotes (2×1) constant vectors and the threshold is denoted by γ, $\gamma > 0$. The $(2 \times 2p)$ parameter matrices are $\Gamma^{(j)} = [\Gamma_1^{(j)}, \ldots, \Gamma_p^{(j)}]$ and ΔX_{t-1} $(2p \times 1)$ is given by

$$\Delta X_{t-1} \quad = \quad \begin{pmatrix} \Delta x_{t-1} \\ \vdots \\ \Delta x_{t-p} \end{pmatrix},$$

for some integer $p > 0$. The vector u_t is assumed to be a independently and normally distributed random sequence, $u_t \sim iid\, N(0_2, \Sigma)$, with non-singular covariance matrix $E[u_t u_t'] = \Sigma$. Note that the model is written as a two-regime model.

Denoting $\kappa = \beta'\mu$, $\beta^* = (\beta',\ -\kappa)'$ and $y_t^* = (y_t',\ 1)'$, one can write the vector error correction model for y_t as

$$\Delta y_t = \begin{cases} \alpha^{(2)}\beta^{*\prime} y_{t-1}^* + \Gamma^{(2)}\Delta Y_{t-1} + u_t, & |\beta^{*\prime} y_{t-1}^*| > \gamma \\ \alpha^{(1)}\beta^{*\prime} y_{t-1}^* + \Gamma^{(1)}\Delta Y_{t-1} + u_t, & |\beta^{*\prime} y_{t-1}^*| \leq \gamma. \end{cases} \tag{1.1}$$

This model includes the classical EQ - TAR model of Balke and Fomby (1997) when $p = 0$. In this model, the error correction term converges towards the equilibrium relationship itself given by $\beta' y_{t-1} - \kappa$. The speed of adjustment, however, is allowed to depend on the magnitude of the equilibrium deviation. This effect may be explained by the presence of transaction or information costs that prevent agents to adjust immediately to equilibrium. In general, this model may allow us to evaluate the "strength" of a postulated equilibrium relationship. A useful restricted version of this model could be

$$\Delta y_t = \begin{cases} \alpha^{(2)} \beta^{*\prime} y^*_{t-1} + \Gamma \Delta Y_{t-1} + u_t, & |\beta^{*\prime} y^*_{t-1}| > \gamma \\ \alpha^{(1)} \beta^{*\prime} y^*_{t-1} + \Gamma \Delta Y_{t-1} + u_t, & |\beta^{*\prime} y^*_{t-1}| \leq \gamma. \end{cases}$$

This model allows only the loading coefficients to switch while restricting the parameters on the lagged differences to be the same across regimes.

One can also think of an extension of the BAND-TAR model proposed by Balke and Fomby (1997). The process x_t takes then the form

$$\Delta x_t = \begin{cases} \alpha^{(2)} (\beta' x_{t-1} - \gamma) + \Gamma^{(2)} \Delta X_{t-1} + u_t, & \beta' x_{t-1} > \gamma \\ \alpha^{(1)} \beta' x_{t-1} + \Gamma^{(1)} \Delta X_{t-1} + u_t, & -\gamma \leq \beta' x_{t-1} \leq \gamma \\ \alpha^{(2)} (\beta' x_{t-1} + \gamma) + \Gamma^{(2)} \Delta X_{t-1} + u_t, & \beta' x_{t-1} < -\gamma. \end{cases}$$

Using the same notation as above, this specification implies the following model for y_t

$$\Delta y_t = \begin{cases} \alpha^{(2)} (\beta^{*\prime} y^*_{t-1} - \gamma) + \Gamma^{(2)} \Delta Y_{t-1} + u_t, & \beta^{*\prime} y^*_{t-1} > \gamma \\ \alpha^{(1)} (\beta^{*\prime} y^*_{t-1}) + \Gamma^{(1)} \Delta Y_{t-1} + u_t, & -\gamma \leq \beta^{*\prime} y^*_{t-1} \leq \gamma \\ \alpha^{(2)} (\beta^{*\prime} y^*_{t-1} + \gamma) + \Gamma^{(2)} \Delta Y_{t-1} + u_t, & \beta^{*\prime} y^*_{t-1} < -\gamma. \end{cases} \tag{1.2}$$

In this model, the error correction term converges towards a band $[-\gamma, \gamma]$ rather than to the equilibrium relation itself, provided that the cointegration residual is large in absolute value. This specification might be a sensible description of the data if, for instance, transaction costs are specified as a proportional loss of the value gained by adjusting to the equilibrium.

The maximum likelihood estimation of the models 1.1 and 1.2 is implemented as proposed by Hansen and Seo (2002). That is, one first concentrates out all parameters except the cointegrating vector and the threshold and then searches for the maximum of the likelihood function using a grid search over values for β and γ. However, the dimension of the grid-search increases by one because (standardizing the first element of β, $\beta_1 = 1$) one has to search over values for β_2, κ and γ. This is computationally more demanding but still feasible. The procedure is explained in more detail in the appendix.

1.3 Application to the U.S. interest rate spread

In economic theory short-term and long-term interest rates are supposed to be linked via the expectation hypothesis that asserts that long-term interest rates roughly represent an average of current and expected short-term interest rates over the life of the long-term bond. Since agents can re-sample a long-term bond by several bonds of a shorter maturity the expected returns of both strategies must be equal up to a risk premium in equilibrium. Denote by r_t and $R_t^{(n)}$ the short-term interest rate and the long-term interest rate, respectively. Then the following relation should hold approximately between a one-period short and a n-period long rate

$$R_t^{(n)} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} E_t(r_{t+i-1}) + \phi$$

for $n = 1, 2, \ldots$ and ϕ denotes the liquidity premium. According to Campbell and Shiller (1987) this economic relation implies a cointegrating relation between the two rates in econometric terms - provided that future expected changes in the short rate are stationary. In contrast, the segmented market theory states that bond markets for different maturities are separated since agents have preferences over different maturities. Thus, according to this theory, bonds with different maturities cannot be seen as substitutes and consequently the interest rate spread should not display a mean reverting behavior. The term structure has been analyzed by various authors using linear cointegration analysis and nonlinear techniques - with different outcomes. In particular, Seo (2003) fits an unrestricted three-regime TVECM arguing that transaction and information costs may imply a non-linear mean reversion in the spread. He finds that the spread is persistent within a non-symmetric band but returns to the equilibrium relation when the spread exceeds the thresholds. I evaluate here the performance of the restricted models 1.1 and 1.2.

The three-month treasury bill (TB3M) and the ten-year constant maturity rate (TN10Y) are used as the short-term interest rate and long-term interest rate, respectively. The monthly data is taken from the Federal Reserve data base and comprises the period 1960:1 to 2004:10. The series are plotted in figure 1.1. The series seem to be non-stationary and display persistent behavior which is confirmed by a visual inspection of the autocorrelations. However, the series show no deterministic upward or downward pattern. Thus, it may indeed be reasonable to exclude a time trend in the following VECM analysis. Furthermore, the impression is that the series may very well cointegrate but one observes that both series drift apart for long periods without any tendency to return to some stable relation.

Linear unit-root tests loose power when the alternative is a nonlinear, stationary process. However, as shown by Balke and Fomby (1997) they still work reasonably well in the case of threshold cointegration. Therefore, the results of standard unit-root tests convey some information. The results are given in table 1.1 for the ADF and the KPSS test for level

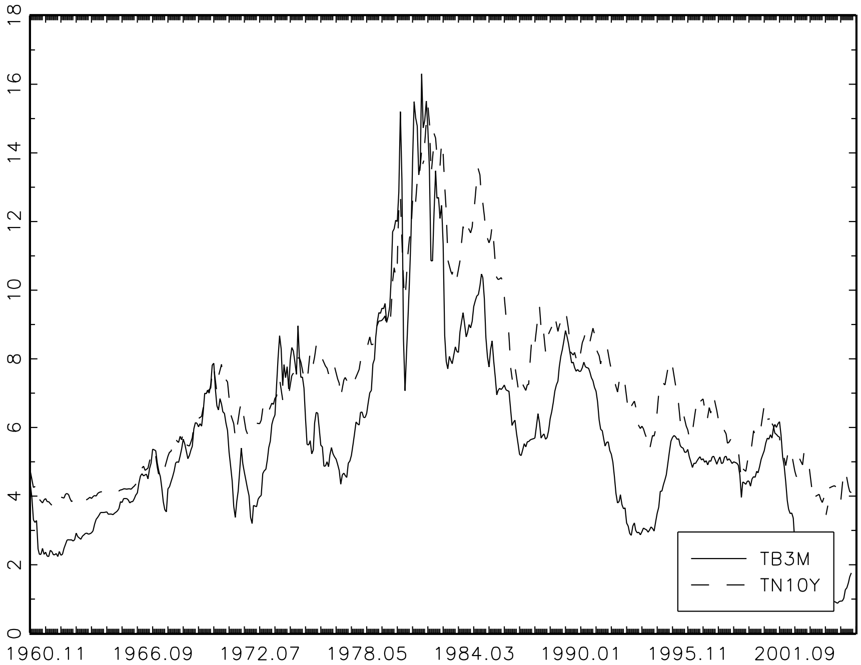

Figure 1.1: Short and long-term interest rates

stationarity (Fuller, 1976; Dickey and Fuller, 1979; Kwiatkowski, Phillips, Schmidt and Shin, 1992). For TB3M the Schwarz criterion suggests a lag length of 13. I also provide the results for an autoregressive model with 6 lags for robustness. For TN10Y the Schwarz criterion and the Hannan-Quinn criterion recommend lag lengths 6 and 2, respectively. The results for both lag lengths are given in the table. For both series, the ADF test cannot reject the null of a unit-root at a 10% level for different lag lengths. For the corresponding differenced series, however, the ADF test rejects the null of a unit root always at the 1 % level. For the KPSS test, Corradi, Swanson and White (2000) have shown that this test is still applicable for many nonlinear specifications of the DGP, including some threshold autoregressive processes. I report the test statistic for lags of 6 and 18 [1]. The KPSS test confirms the results of the ADF test. The differenced series, however, are stationary according to both tests. Based on these linear tests one may therefore conclude that both series are $I(1)$.

To confirm these results I test the null of a unit-root also within the class of some common threshold models using the approach by Caner and Hansen (2001). They consider two-regime TAR models with different threshold variables and suggest a bootstrap method to calculate

[1] The lag length is chosen according to $l_q = q(T/100)^{1/4}$ with $q = 4, 12$ as proposed by Kwiatkowski et al. (1992).

the p-values for different statistics. The framework is the following univariate TAR model

$$\Delta y_t \quad = \quad \begin{cases} \rho_2 y_{t-1} + \mu_t^{(2)} + \sum_{i=1}^p \alpha_i^{(2)} \Delta y_{t-i}, & z_{t-1} \geq \gamma \\ \rho_1 y_{t-1} + \mu_t^{(1)} + \sum_{i=1}^p \alpha_i^{(1)} \Delta y_{t-i}, & z_{t-1} < \gamma, \end{cases}$$

where the threshold variable, z_{t-1}, is assumed be a stationary random variable under the null as well as under the alternative hypothesis. They assume $|\sum \alpha_i^{(1)}| < 1$, $|\sum \alpha_i^{(2)}| < 1$ such that the process Δy_t is stationary and ergodic. Note that by this assumption a non-rejection of the unit-root hypothesis implies that the series are $I(1)$. They evaluate tests for $\mathcal{H}_0 : \rho_1 = \rho_2 = 0$ versus $\mathcal{H}_1 : \rho_1 \neq 0$ or $\rho_2 \neq 0$ $(R2t)$, $\mathcal{H}_2 : \rho_1 < 0$ or $\rho_2 < 0$ $(R1t)$ and provide the negative t-ratios $(t1, t2)$ in order to test for $\mathcal{H}_3 : \rho_1 < 0$, $\mathcal{H}_4 : \rho_2 < 0$. The threshold variable is allowed to be either a long difference in y_t, $z_{t-1} = y_{t-1} - y_{t-1-d}$, (ld) or a lagged change, $z_{t-1} = y_{t-d} - y_{t-d-1}$, (lc) where d denotes the delay parameter. The autoregressive lags were chosen according to the linear information criteria as discussed above for the ADF statistic. The delay that minimized the residual sum of squares was chosen for the threshold variable (Caner and Hansen, 2001). The p-values obtained by their bootstrap methodology are given in table 1.2 for different specifications and for the various alternative hypothesis. The asymptotic p-values (not reported) are in general more conservative. The null hypothesis cannot be rejected in most cases at the 5 % level against any of the different alternative hypothesis, aside from one case (row 9). The overall impression based on these nonlinear statistics is therefore that both series are indeed $I(1)$.[2]

The same limitations for the standard unit-root tests also apply to the linear cointegration tests when the true processes are nonlinear. Though they remain useful as well (Balke and Fomby, 1997). The results are given in table 1.3 for the Johansen and the Saikkonen and Lütkepohl test (Johansen, 1995; Saikkonen and Lütkepohl, 2000a,b). The lag length has been chosen according to the Schwarz criterion that recommends a lag length of 3 in levels and only a constant is included. We also tabulate the results for 6 levels for robustness. The results in table 1.3 are very uniform: The Johansen test and the Saikkonen and Lütkepohl test indicate clearly that the series are cointegrated with rank one.

The Schwarz criterion recommends a lag length of three for a linear VAR. After fitting the linear VECM, I test the null of linearity against nonlinearity using the test of Tsay (1998). Tsay (1998) tests for nonlinearity in the framework of a VAR. Although the test was originally constructed to test linearity against threshold nonlinearity it is essentially nonparametric in that it tests linearity against a class of nonlinear models. It makes use of an "arranged autoregression", that is, the data is ordered according to the threshold variable, preserving the dynamic relation between y_t and its lags. Then the predictive residuals from a standard OLS regression are computed recursively starting from some initial observation.

[2] Furthermore, Caner and Hansen (2001) also develop a Wald test for linearity versus threshold non-linearity. The calculated bootstrap test statistics all reject the linear model strongly in favor of the threshold models.

The test makes use of the simple idea that under the null hypothesis of linearity the predictive residuals should approach white noise and consequently be uncorrelated with the threshold variable. On the other hand, if y_t follows a threshold autoregression than the residuals will be correlated with the threshold variable. For the test of Tsay (1998) the threshold variable is taken as given. In this case, however, the threshold variable, the error correction term, is unknown and thus has to be estimated under the null of linearity. However, it is not clear to what extent the test is still valid when the threshold variable is estimated. The test rejects linearity clearly at the 1% level with a test statistic of 68.86. This result has to be interpreted in favor of some general nonlinear alternative. In sum, there is evidence favoring (threshold) nonlinearity.

Also for the TVECMs 1.1 and 1.2 I use a lag length of three in levels. Since a distribution theory for the case of an unknown cointegrating vector has yet to be developed, I select the model comparing the AIC criterion which yields a minimal value for the model given in (1.1). The resulting error correction term and the residuals for the fitted TVECM are plotted in figure 1.2. The corresponding estimated coefficients are given in table 1.4, where the order of the variables in the model is $y_t = (\mathrm{TB3M}_t,\ \mathrm{TN10Y}_t)'$. The standard deviations of the parameters are given in parenthesis. However, since there is no distribution theory for an unknown cointegration vector, one has to interpret them with some caution although results for univariate TAR processes lead to the suspicion that the distribution converges to a multivariate normal distribution, as if β and γ were known. With regard to the cointegrating vector, one observes that the coefficient on the long-term interest rate is remarkably close to -1, while one may interpret the constant as a mean risk premium of about 1.5 percentage points. The estimated threshold is noticeably high at 1.9. This result implies a rather large band around the equilibrium in which the two series drift apart freely. The estimated coefficients are consistent with an adjustment cost interpretation. As long as the error correction term is in the estimated band the loading coefficients are relatively small compared to those in the outer regime and insignificant at conventional levels (if the above mentioned conjecture about the asymptotic distribution is right). In sum, this would mean that the theoretical term structure relation does not help very much in predicting interest rates unless the spread is very high in absolute terms.

I compare the forecast performance of the linear VECM and the TVECM 1.1 by computing one-step-ahead predictions and estimate the root mean square error (RMSE), the mean and median absolute error (MAE and MEAE) for the short-term interest rate. This is done by estimating both models recursively and comparing the resulting forecast with the true value. I compute the forecasts for the last 150 observations in the sample. Table 1.5 shows the point estimates of the measures for the TVECM model 1.1 relative to the ones obtained from the linear VECM. The picture is at best mixed in that the predictions from the TVECM are worse in terms of the first two measures but better in terms of the last measure.

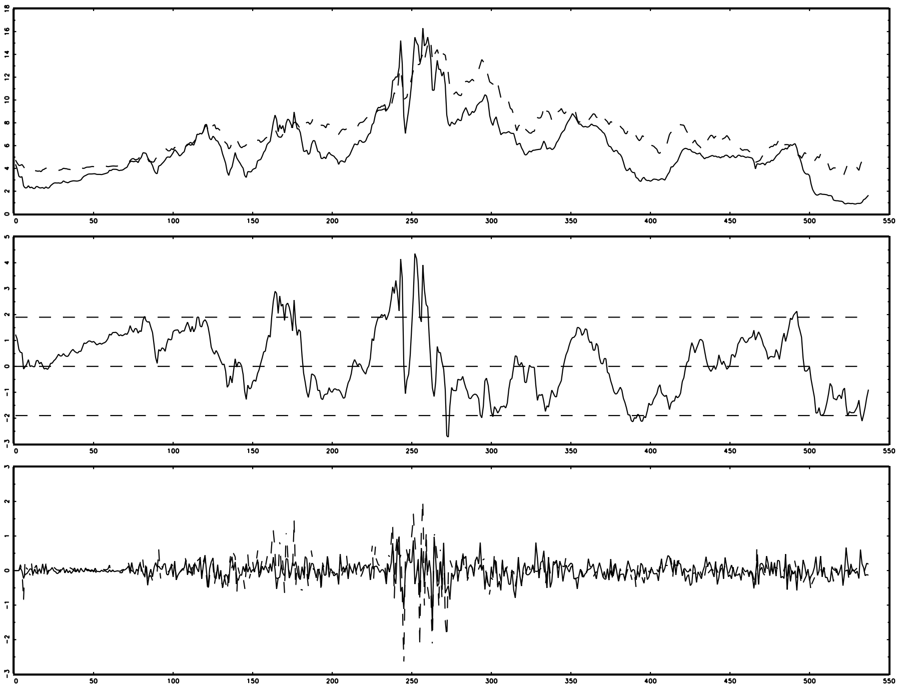

Figure 1.2: Data, cointegrating term, residuals

The resulting residuals from the TVECM give a slightly better impression in that they are less skewed and have a lower kurtosis than the residuals obtained from a linear VECM (not shown). Furthermore, the number of outliers is reduced. However, the residuals do not look very well behaved. We still lack specification tests for threshold cointegrated models, thus we are not able to present some reliable diagnostics such as an LM test for autocorrelation or a test for structural stability. The latter would be particularly useful in order to distinguish nonlinearity and structural breaks and would allow for a full model comparison. Though when applying linear checks as provisional tools we find that, e.g., the ARCH-LM test indicates that there are ARCH effects present while the Portmanteau test indicates that there is some autocorrelation left in the residuals. Another nonlinear model might actually be a better description of the underlying DGP of the investigated series. In particular, it seems reasonable to model also the dependence in the conditional variance.

1.4 Conclusion

In this note I reviewed symmetric threshold models that are extensions of the BAND-TAR and EQ-TAR models of Balke and Fomby (1997). These models allow to control for the deterministic components in the data and are applied to a bivariate U.S. interest rate series to study the dynamics towards the term structure implied by the expectation hypothesis. I also review the relevant specification tools for threshold models.

The employed test for linearity provides evidence in favor of the threshold models. The fitted model would confirm the view that in a band around the equilibrium relation there is little mean reversion in the spread. Only when deviations from the equilibrium become large one observes that the interest rates are attracted towards the cointegration relationship.

There are many open questions related to the specification and estimation of threshold cointegrated systems. The asymptotic theory for the maximum likelihood estimates in the threshold cointegration model has yet to be developed and extended to cover more general processes than the models considered in this note. The provision of valid specification and model checking tools would be very useful in order to appreciate the potential merits of threshold models better. It would also be useful to find ways to restrict the intercepts and trends in the framework of more general threshold cointegration models. These restrictions are likely to be quite complicated and it is not obvious how to impose them. These topics, however, are left for future research.

Appendix

1.A Maximum Likelihood Estimation of TVECMs

As mentioned above the used implementation of maximum likelihood estimation was originally proposed by Hansen and Seo (2002). I exemplify their methodology using the model in (1.1):

$$\Delta y_t \quad = \quad d_{1t}(\alpha^{(1)}\beta^{*\prime}y^*_{t-1} + \Gamma^{(1)}\Delta Y_{t-1}) + d_{2t}(\alpha^{(2)}\beta^{*\prime}y^*_{t-1} + \Gamma^{(2)}\Delta Y_{t-1}) + u_t,$$

where $d_{1t} = 1_{(|\beta^{*\prime}y^*_{t-1}| \leq \gamma)}$ and $d_{2t} = 1_{(|\beta^{*\prime}y^*_{t-1}| > \gamma)}$ are indicator functions. To simplify notation let us write $A_j' = [\alpha^{(j)} : \Gamma^{(j)}]$ and $Z_{t-1} = (\beta^{*\prime}y^*_{t-1} : \Delta Y_t')'$. Assuming that the errors u_t are *iid* normally distributed the joint log-likelihood can be formulated as

$$\mathcal{L}_T(A_1, A_2, \Sigma, \beta^*, \gamma) = -\frac{2\,T}{2}\log(2\pi) - \frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{t=1}^{T} u_t'\Sigma^{-1}u_t,$$

with

$$u_t(A_1, A_2, \Sigma, \beta^*, \gamma) = \Delta y_t - A_1' Z_{t-1} d_{1t} - A_2' Z_{t-1} d_{2t}.$$

The algorithm can be described in four steps:

1. Obtain an initial linear estimate of the cointegrating vector, $\hat{\beta}_0$, and the corresponding error correction term $\hat{z}_{0,t} = \hat{\beta}_0' y_t$. Normalize the cointegrating vector such that $\hat{\beta}_0 = (1 \ : \ \hat{\beta}_{0,2} \ : \ \hat{\kappa}_0)'$.

2. Set up an evenly spaced grid on an interval $[\beta_L, \beta_U]$ centered around $\hat{\beta}_{0,2}$, a grid on $[\kappa_L, \kappa_U]$ around $\hat{\kappa}_0$ and another grid on the empirical support of $\hat{z}_{0,t}$, $[\gamma_L, \gamma_U]$, subject to the constraint that a minimum fraction of observations is in each regime for all $\gamma \in [\gamma_L, \gamma_U]$.

3. Compute for each triple $(\beta_2, \kappa, \gamma) \in [\beta_L, \beta_U] \times [\kappa_L, \kappa_U] \times [\gamma_L, \gamma_U]$ the value of the log-likelihood function by computing the conditional maximum likelihood estimates $\hat{A}_1(\beta_2, \kappa, \gamma), \hat{A}_2(\beta_2, \kappa, \gamma), \widehat{\Sigma}(\beta_2, \kappa, \gamma)$. These are just the usual LS estimates yielding the concentrated likelihood

$$\mathcal{L}_T(\beta_2, \kappa, \gamma) = -\frac{2\,T}{2}\log(2\pi) - \frac{T\;2}{2} - \frac{T}{2}\log|\widehat{\Sigma}(\beta_2, \kappa, \gamma)| \qquad (1.3)$$

4. The maximum likelihood estimate $(\hat{\beta}_2, \hat{\kappa}, \hat{\gamma})$ is the triple $(\beta_2, \kappa, \gamma)$ that maximizes (1.3) and the estimates for the remaining parameters are $\hat{A}_1(\hat{\beta}_2, \hat{\kappa}, \hat{\gamma}), \hat{A}_2(\hat{\beta}_2, \hat{\kappa}, \hat{\gamma}), \widehat{\Sigma}(\hat{\beta}_2, \hat{\kappa}, \hat{\gamma})$.

1.B Tables

Table 1.1: Linear Unit-Root Tests

Series	Lags	Det.	ADF	KPSS
TB3M_t	6	c	-1.65	1.23***/0.49**
	13	c	-2.31	
ΔTB3M_t	5	-	-12.14***	0.09/0.09
	12	-	-5.97***	
TN10Y_t	2	c	-1.46	1.66***/0.64**
	6	c	-1.59	
ΔTN10Y_t	1	-	-17.17***	0.21/0.20
	5	-	-9.49***	

Linear unit-root tests for the levels and first differences of the short-term and long-term interest rates (TB3M and TN10Y). The number of lags refers to the ADF test. Det. indicates the deterministic terms included, where c means that a constant is included. The notation */**/*** indicates rejection of $\mathcal{H}_0$ at the 10/5/1% level.

Table 1.2: Nonlinear Unit-Root Tests

	p	d	z_{t-1}	$R2T$	$R1T$	$t1$	$t2$
TB3M_t	6	2	ld	0.20	0.21	0.08	0.94
TB3M_t	13	7	ld	0.19	0.17	0.52	0.10
TB3M_t	6	6	lc	0.87	0.85	0.87	0.48
TB3M_t	13	6	lc	0.66	0.94	0.67	0.99
TN10Y_t	2	2	ld	0.36	0.52	0.97	0.25
TN10Y_t	6	5	ld	0.34	0.71	0.33	0.99
TN10Y_t	2	2	lc	0.09	0.07	0.03	0.79
TN10Y_t	6	2	lc	0.89	0.88	0.52	0.89

Calculated bootstrap p-values for $\mathcal{H}_0 : \rho_1 = \rho_2 = 0$ versus different alternative hypothesis (see text). p and d denote the used lag length and the delay of the threshold variable, respectively, and lc and ld indicate whether the threshold variable is a lagged changed or a long difference of the dependent variable.

Table 1.3: Cointegration Tests

Test	No. of lags (levels)	Null hypothesis	Test value	p-value
Johansen	3	$r = 0$	24.33	0.01
		$r = 1$	2.86	0.61
	6	$r = 0$	24.53	0.01
		$r = 1$	4.08	0.41
S & L	3	$r = 0$	17.15	0.01
		$r = 1$	1.66	0.23
	6	$r = 0$	17.52	0.01
		$r = 1$	2.18	0.16

Johansen and Saikkonen and Lüktepohl (S & L) cointegration tests for different null hypotheses about the cointegrating rank, r. Deterministic terms: constant.

Table 1.4: Estimation results

$\hat{\beta}' = (1, \ -0.99, \ 1.53)$

$\hat{\gamma} = 1.89$

	$\hat{\alpha}^{(j)}$	$\hat{\Gamma}_1^{(j)}$		$\hat{\Gamma}_2^{(j)}$	
outer regime	-0.103	-0.110	1.524	0.211	-0.902
	(0.026)	(0.085)	(0.164)	(0.087)	(0.177)
	0.032	-0.212	0.810	0.218	-0.559
	(0.017)	(0.056)	(0.110)	(0.058)	(0.118)
inner regime	0.015	0.530	0.041	-0.250	-0.008
	(0.017)	(0.061)	(0.083)	(0.060)	(0.084)
	0.014	0.074	0.297	-0.051	-0.190
	(0.012)	(0.040)	(0.056)	(0.040)	(0.056)

Estimated coefficient matrices for model 1.1 dependent on the regime. The numbers in parenthesis are standard errors of the corresponding single coefficients above.

Table 1.5: Relative Forecasting Performance

	RMSE	MAE	MEAE
TVECM /VECM	1.23	1.10	0.91

Root mean square error (RMSE), mean absolute and median absolute error (MAE and MEAE) for the predictions of TB3M_t obtained from model 1.1 relative to the linear benchmark VECM.

Bibliography

Bacon, D. W. and Watts, D. G. (1971), 'Estimating the transition between two intersecting straight lines', *Biometrica* **58**(3), 525–534.

Balke, N. S. and Fomby, T. B. (1997), 'Threshold cointegration', *International Economic Review* **38**(3), 627–645.

Baum, C. F. and Karasulu, M. (1998), 'Modelling federal reserve discount policy', *Computational Economics* **11**, 53–70.

Campbell, J. Y. and Shiller, R. J. (1987), 'Cointegration and tests of present value models', *Journal of Political Economy* **95**, 1062–1088.

Caner, M. and Hansen, B. E. (2001), 'Threshold autoregression with a unit-root', *Econometrica* **69**(6), 1555–1596.

Chan, K. S. and Tong, H. (1986), 'On estimating thresholds in autoregressive models', *Journal of Time Series Analysis* **7**, 179–190.

Corradi, V., Swanson, N. R. and White, H. (2000), 'Testing for stationarity-ergodicity and for commovements between nonlinear discrete time markov processes', *Journal of Econometrics* **96**, 39–73.

Dickey, D. A. and Fuller, W. A. (1979), 'Estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association* **74**, 427–431.

Fuller, W. A. (1976), *Introduction to Statistical Time Series*, John Wiley & Sons, New York.

Goldfeld, S. M. and Quandt, R. E. (1973), 'A markov model for switching regressions', *Journal of Econometrics* **1**, 3–16.

Hamilton, J. D. (1989), 'A new approach to the economic analysis of time series', *Econometrica* **57**(2), 357–384.

Hansen, B. E. (1999), 'Testing for linearity', *Journal of Economic Surveys* **13**(5), 551–576.

Hansen, B. E. (2000), 'Sample splitting and threshold estimation', *Econometrica* **68**(3), 575–603.

Hansen, B. E. and Seo, B. (2002), 'Testing for two-regime threshold cointegration in vector error-correction models', *Journal of Econometrics* **110**, 293–318.

Johansen, S. (1995), *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford.

Kuan, C.-M. and White, H. (1994), 'Artificial neutral networks: An econometric perspective', *Econometric Reviews* **13**(1), 1–91.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. (1992), 'Testing the null of stationarity against the alternative of a unit root: How sure are we that the economic time series have a unit root?', *Journal of Econometrics* **54**, 159–178.

Lo, M. E. and Zivot, E. (2001), 'Threshold cointegration and the law of one price', *Macroeconomic Dynamics* **5**, 533–576.

Obstfeld, M. and Taylor, A. M. (1997), 'Nonlinear aspects of goods market arbitrage and adjustment: Heckscher's commodity points revisited', *Journal of the Japanese and International Economies* **11**, 441–479.

Saikkonen, P. and Lütkepohl, H. (2000a), 'Testing for the cointegrating rank of a var process with an intercept', *Econometric Theory* **16**, 373–406.

Saikkonen, P. and Lütkepohl, H. (2000b), 'Trend adjustment prior to testing for the cointegrating rank of a vector autoregressive process', *Journal of Time Series Analysis* **21**, 435–456.

Seo, B. (2003), 'Nonlinear mean reversion in the term structure of interest rates', *Journal of Economic Dynamics and Control* **27**, 2243–2265.

Taylor, A. M. (2001), 'Potential pitfalls for the purchasing-power-parity puzzle? Sampling and specification biases in mean-reversion tests of the law of one price', *Econometrica* **69**, 473–498.

Teräsvirta, T. and Anderson, H. M. (1992), 'Characterising nonlinearities in busines cycles using smooth transition autoregressive models', *Journal of Applied Econometrics* **7**, S119–S136.

Tong, H. (1978), On a threshold model, *in* C. H. Chen, ed., 'Pattern Recognition and Signal Processing', Sijhoff & Noordhoff, Amsterdam.

Tsay, R. S. (1998), 'Testing and modeling multivariate threshold models', *Journal of the American Statistical Association* **93**(443), 1188–1202.

Chapter 2

Business Cycle Analysis and VARMA Models

Jointly written with Karel Mertens

2.1 Introduction

Structural vector autoregressions (SVARs) are a widely used tool in empirical macroeconomics, particularly for the evaluation of dynamic stochastic general equilibrium (DSGE) models.[1] The results from SVARs are often viewed as stylized facts that economic models should replicate. However, there is some debate whether SVARs can in practice discriminate between competing DSGE models and whether their sampling properties are good enough to justify their popularity in applied macroeconomics. In response to a seminal paper by Gali (1999), the discussion has focused on the impact of technology shocks on hours worked, identified using restrictions on the long-run impact matrix of the structural errors. Chari, Kehoe and McGrattan (2005) and Christiano, Eichenbaum and Vigfusson (2006) investigate the properties of the estimators based on SVARs by simulating an artificial data generating process (DGP) derived from a prototype real business cycle (RBC) model and by comparing true with estimated impulse responses.

According to Chari et al. (2005), long-run identified SVARs fail dramatically for both a level and difference specification of hours worked. Even with a correct specification of the integration properties of the series, the SVAR overestimates in most cases the impact of technology on labor and the estimates display high variability. However, Christiano et al. (2006) argue that the parametrization chosen by Chari et al. (2005) is not very realistic. With their preferred parametrization, Christiano et al. (2006) find that both long-run and short-run identification schemes display only small biases and argue that, on average, the

[1] Examples in the literature are, among many others, Blanchard and Quah (1989), as well as King, Plosser, Stock and Watson (1991), Christiano and Eichenbaum (1992) and Gali (1999).

confidence intervals produced by SVARs correctly reflect the degree of sampling uncertainty.[2] Nevertheless, they also find that the estimates obtained via a long-run identification scheme are very imprecise. These results have been further confirmed by Erceg, Guerrieri and Gust (2005). In the end, with long-run restrictions, it is often very difficult to even make a correct inference about the *sign* of the structural impulse responses. The question is therefore if one should use this type of identification scheme at all. However, long-run identification is attractive from a theoretical point of view, since it requires much weaker assumptions than short-run identification and is in any case a useful additional tool for model evaluation.

The failure of finite-order SVARs is sometimes attributed to the fact that they are only approximations to infinite-order VAR processes or to the possibility that there does not exist a VAR representation at all. For example, Cooley and Dwyer (1998) give an example of an economic model that implies a vector autoregressive moving-average (VARMA) representation of the data series and state: "*While VARMA models involve additional estimation and identification issues, these complications do not justify* systematically *ignoring these moving average components, as in the SVAR approach*". As further shown by Fernández-Villaverde, Rubio-Ramírez and Sargent (2005), DSGE models generally imply a state space system that has a VARMA and eventually an infinite VAR representation. Fernández-Villaverde et al. (2005) propose the inclusion of moving-average terms if the DSGE model at hand does not permit an infinite VAR representation. Christiano et al. (2006) state that "*Given our data generating processes, the true VAR of the data has infinite lags. However, the econometrician can only use a finite number of lags in the estimation procedure. The resulting specification error is the reason why in some of our examples the sum of VAR coefficients is difficult to estimate accurately*".

This paper explores the possible advantages of structural VARMA and state space models that capture the full structure of the time series representation implied by DSGE models, while imposing minimal theoretical assumptions. We investigate whether estimators based on these alternative models can outperform SVARs in finite samples.[3] This question is important for several reasons. First, it is useful to find out to what extent the poor performance of SVARs in these simulation studies is due to the omission of moving-average components. Second, whether estimators based on alternative representations of the same DGP have good sampling properties is interesting in itself. Employing these alternatives enables researchers to quantify the robustness of their results by comparing different estimates.

In order to assess whether the inclusion of a moving-average component leads to impor-

[2] In addition, Christiano et al. (2006) find that short-run identification schemes work much better compared to identification via long-run restrictions.

[3] McGrattan (2006) is closely related to our paper. In a similar setting, McGrattan (2006) also investigates whether state space or VARMA models with minimal structural assumptions can uncover statistics of interest. Her work focusses on different business cycle statistics, while we are exclusively concerned with classical structural estimation.

tant improvements, we stick to the research design of Chari et al. (2005) and Christiano et al. (2006): We simulate DSGE models and fit different reduced form models to recover the structural shocks using the same long-run identification strategy. We then compare the performance of the models by focusing on the estimated contemporaneous impact to a technology shock. We employ a variety of estimation algorithms for the VARMA models, and both a prediction error method and a subspace algorithm for the state space models. One of the findings is that one can indeed perform better by taking the full structure of the DGP into account: While the algorithms for VARMA models and the prediction error method do not perform significantly better (and sometimes worse), the subspace algorithm for state space models consistently outperforms SVARs in terms of mean squared error. Unfortunately, we also find that even these alternative estimators are highly variable and are therefore not necessarily much more informative for discriminating between different DSGE models. One of the implications is that SVARs do not perform poorly in these simulation studies because they are only finite-order approximations. Given the properties of the data generating process, the disappointing performance of SVARs is most likely due to the fact that the long-run identification approach is inappropriate with small samples.

The rest of the paper is organized as follows. In section 2 we present the RBC model used by Chari et al. (2005) and Christiano et al. (2006) that serves as the basis for our Monte Carlo simulations. In section 3 we discuss the different statistical representations of the observed data series. In section 4 we present the specification and estimation procedures and the results from the Monte Carlo simulations. Section 5 concludes.

2.2 The Data Generating Process

The DGP for the simulations is based on a simple RBC model taken from Chari et al. (2005). In the model, a technology shock is the only shock that affects labor productivity in the long-run, which is the crucial identifying assumption made by Gali (1999) to assess the role of technology shocks in the business cycle.

Households choose infinite sequences, $\{C_t, L_t, K_{t+1}\}_{t=0}^{\infty}$, of per capita consumption, labor and capital to maximize expected lifetime utility

$$E_0 \sum_{t=0}^{\infty} [\beta(1+\gamma)]^t \left[\log C_t + \psi \frac{(1-L_t)^{1-\sigma}-1}{1-\sigma} \right],$$

given an initial capital stock K_0, and subject to a set of budget constraints given by

$$C_t + (1+\tau_x)\left((1+\gamma)K_{t+1} - (1-\delta)K_t\right) \leq (1-\tau_{lt})w_t L_t + r_t K_t + T_t,$$

for $t = 0, 1, 2, ...$, where w_t is the wage, r_t is the rental rate of capital, T_t are lump-sum government transfers and τ_{lt} is an exogenous labor tax. The parameters include the discount factor $\beta \in (0,1)$, the labor supply parameters, $\psi > 0$ and $\sigma > 0$, the deprecation rate $\delta \in (0,1)$, the population growth rate $\gamma > 0$ and a constant investment tax τ_x. The production technology is

$$Y_t = K_t^{\alpha}(X_t L_t)^{1-\alpha},$$

where X_t reflects labor-augmenting technological progress and $\alpha \in (0,1)$ is the capital income share. Competitive firms maximize $Y_t - w_t L_t - r_t K_t$. Finally, the resource constraint is $Y_t \geq C_t + (1+\gamma)K_{t+1} - (1-\delta)K_t$.

The model contains two exogenous shocks, a technology shock and a tax shock, which follow the stochastic processes

$$\log X_{t+1} = \mu + \log X_t + \sigma_x \epsilon_{x,t+1},$$
$$\tau_{lt+1} = (1-\rho)\bar{\tau}_l + \rho\tau_{lt} + \sigma_l \epsilon_{l,t+1},$$

where $\epsilon_{x,t}$ and $\epsilon_{l,t}$ are independent random variables with mean zero and unit standard deviation and $\sigma_x > 0$ and $\sigma_l > 0$ are scalars. $\mu > 0$ is the mean growth rate of technology, $\bar{\tau}_l > 0$ is the mean labor tax and $\rho \in (0,1)$ measures the persistence of the tax process. Hence, the model has two independent shocks: a unit root process in technology and a stationary AR(1) process in the labor tax.

2.3 Statistical Representations

Fernández-Villaverde et al. (2005) show how the solution of a detrended, log-linearized DSGE model leads to different statistical representations of the model-generated data. This section presents several alternative ways to write down a statistical model for the bivariate, stationary time series

$$\mathbf{y}_t = \begin{bmatrix} \Delta \log(Y_t/L_t) \\ \log(L_t) \end{bmatrix}.$$

Labor productivity growth $\Delta \log(Y_t/L_t)$ and hours worked $\log(L_t)$ are also the series analyzed by Gali (1999), as well as Chari et al. (2005) and Christiano et al. (2006). The appendix provides more detail on the derivations. Given the log-linearized solution of the RBC model of the previous section, we can write down the law of motion of the logs

$$\begin{aligned} \log k_{t+1} &= \phi_1 + \phi_{11}\log k_t - \phi_{11}\log x_t + \phi_{12}\tau_t, \\ \log y_t - \log L_t &= \phi_2 + \phi_{21}\log k_t - \phi_{21}\log x_t + \phi_{22}\tau_t, \\ \log L_t &= \phi_3 + \phi_{31}\log k_t - \phi_{31}\log x_t + \phi_{32}\tau_t, \end{aligned}$$

where $k_t = K_t/X_{t+1}$ and $y_t = Y_t/X_t$ are capital and output detrended with the unit-root shock and the ϕ's are the coefficients of the calculated policy rules. Following Fernández-Villaverde et al. (2005) the system can be written in state space form. The state transition equation is

$$\begin{aligned}\begin{bmatrix}\log k_{t+1}\\ \tau_t\end{bmatrix} &= K_1 + A\begin{bmatrix}\log k_t\\ \tau_{t-1}\end{bmatrix} + B\begin{bmatrix}\epsilon_{x,t}\\ \epsilon_{lt}\end{bmatrix},\\ \mathbf{x}_{t+1} &= K_1 + A\mathbf{x}_t + B\epsilon_t,\end{aligned}$$

and the observation equation is

$$\begin{aligned}\begin{bmatrix}\Delta\log(Y_t/L_t)\\ \log L_t\end{bmatrix} &= K_2 + C\begin{bmatrix}\log k_t\\ \tau_{t-1}\end{bmatrix} + D\begin{bmatrix}\epsilon_{x,t}\\ \epsilon_{lt}\end{bmatrix},\\ \mathbf{y}_t &= K_2 + C\mathbf{x}_t + D\epsilon_t,\end{aligned}$$

where K_1, A, B, K_2, C and D are constant matrices that depend on the coefficients of the policy rules and therefore on the "deep" parameters of the model. The state vector is given by $\mathbf{x}_t = [\log k_t,\ \tau_{t-1}]'$ and the noise vector is $\epsilon_t = [\epsilon_{xt},\ \epsilon_{lt}]'$. Note that the system has a state vector of dimension two with the logarithm of detrended capital and the tax rate shock as state components.

The above state space system is still a structural model, since the formulation contains the non-observable state vector and the structural errors. We now show different representations of the system for $\mathbf{y}_t$, which can be estimated in practice. Given certain invertibility conditions on the system matrices, A, B, C, D, there is an **infinite VAR representation**:

$$\mathbf{y}_t = K_3 + C\left(I - (A - BD^{-1}C)L\right)^{-1} BD^{-1}\mathbf{y}_{t-1} + D\epsilon_t, \tag{2.1}$$

or

$$\mathbf{y}_t = K_3 + \sum_{i=1}^{\infty}\Pi_i\mathbf{y}_{t-i} + u_t,$$

where K_3 and $\Pi_i, i = 1, 2, \ldots$ are constant coefficient matrices, L denotes the lag operator, I denotes an identity matrix of suitable dimensions, $u_t = D\epsilon_t$ and $u_t \sim iid\,N(0, \Sigma_u),\ \Sigma_u = DD'$, where Σ_u is the covariance matrix of u_t. Note that a condition for the existence of an infinite VAR representation is that the eigenvalues of $(A - BD^{-1}C)$ are strictly less than one in modulus. In practice, it is only possible to approximate this structure by a finite-order VAR. Alternatively, the system can be written as a **state space model** in "innovations form":

$$\begin{aligned}\mathbf{x}_{t+1} &= K_1 + A\mathbf{x}_t + Ku_t, \\ \mathbf{y}_t &= K_2 + C\mathbf{x}_t + u_t,\end{aligned} \tag{2.2}$$

where the innovation, u_t, is defined as above and $K = BD^{-1}$. In contrast to the VAR representation in (2.1), it is possible to estimate (2.4) exactly.
Finally, the underlying DGP can be represented by a **VARMA(1,1) representation**:

$$\begin{aligned} \mathbf{y}_t &= K_4 + CAC^{-1}\mathbf{y}_{t-1} + \left(D + (CB - CAC^{-1}D)L\right)\epsilon_t, \\ \mathbf{y}_t &= K_4 + A_1\mathbf{y}_{t-1} + u_t + M_1 u_{t-1}, \end{aligned} \tag{2.3}$$

where the last equation defines A_1, M_1 and u_t is defined as above. As with the state space representation, the VARMA(1,1) representation can also be estimated exactly.

Given the conditions stated in Fernández-Villaverde et al. (2005), all three representations are algebraically equivalent. That is, given the same input sequence $\{\epsilon_t\}$, they produce the same output sequence $\{\mathbf{y}_t\}$. The representations are however not statistically equivalent: the properties of estimators and tests depend on the chosen statistical representation. It should be emphasized that we are always interested in the same process and ultimately in the estimation of the same coefficients, i.e. those associated with the first-period response of $\mathbf{y}_t$ to a unit shock in $\epsilon_{x,t}$ to the technology process. However, the different representations give rise to different estimation algorithms and therefore our study can be regarded as a comparison of different algorithms to estimate the same linear system.

2.4 The Monte Carlo Experiment

2.4.1 Monte Carlo Design and Econometric Techniques

To investigate the properties of the various estimators, we simulate 1000 samples of the vector series $\mathbf{y}_t$ in linearized form and transform log-deviations to values in log-levels. As in the previous Monte Carlo studies, the sample size is 180 quarters. We use two different sets of parameter values: The first is due to Chari et al. (2005) and is referred to as the CKM-specification, while the second is the one used by Christiano et al. (2006) and is labeled the KP-specification, referring to estimates obtained by Prescott (1986).[4] The specific parameter values are given in table 2.1 for the CKM and KP benchmark specifications. Christiano et al. (2006) show that the key difference between the specifications is the implied fraction of the variability in hours worked that is due to technology shocks. Table 2.1 also provides the eigenvalues of the autoregressive and moving-average matrices of the corresponding VARMA representations, together with the eigenvalues of the Kalman gain K. In terms of these values, the time series properties are very similar and indicate why the estimation of both systems could be difficult. Note that the moving-average part is not of full rank and the associated

[4] Both parameterizations are obtained by maximum likelihood estimation of the theoretical model, using time series on productivity and hours worked in the US. However, because of differences in approach, both papers obtain different estimates.

eigenvalue is close to unity in modulus. Also, the eigenvalues of the autoregressive part are close to one and close to the eigenvalue of the moving-average part in modulus. The fact that one eigenvalue of the moving-average part is close to one eigenvalue of the autoregressive part could imply that the VARMA(1,1) representation is close to being not identified (Klein, Mélard and Spreij, 2004).

To check the robustness of our results, we also consider variations of the benchmark models. As in Christiano et al. (2006), we consider different values for the preference parameter σ and the standard deviation of the labor tax, σ_l. These variations change the fraction of the business cycle variability that is due to technology shocks. The different values for σ are reported in table 2.2. For the CKM specification, we also consider cases where σ_l assumes a fraction of the original benchmark value.

Turning to the issue of identification, consider the following infinite moving-average representation of $\mathbf{y}_t$ in terms of u_t:

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \Phi_{u,i} u_{t-i} = \Phi_u(L) u_t,$$

where we abstract from the intercept term and $\Phi_u(L)$ is a lag polynomial, $\Phi_u(L) = \sum_{i=0}^{\infty} \Phi_{u,i} L^i$. Analogously, we can represent $\mathbf{y}_t$ in terms of the structural errors using the relation $u_t = D\epsilon_t$:

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \Phi_{u,i} D \epsilon_{t-i} = \Phi_\epsilon(L) \epsilon_t,$$

where $\Phi_\epsilon(L) = \sum_{i=0}^{\infty} \Phi_{u,i} D L^i$. The former lag polynomial, evaluated at one,

$$\Phi_u(1) = I + \Phi_{u,1} + \Phi_{u,2} + \dots$$

is the long-run impact matrix of the reduced form error u_t. Note that the existence of this infinite sum depends on the stationarity of the series. If the stationarity requirement is violated or "nearly" violated, then the long-run identification scheme is not valid or may face difficulties as can be seen from the following discussion. Also note that the matrix D defined in section 2.3 gives the first-period impact of a unit shock in ϵ_t. Using the above relations, we know that $\Phi_\epsilon(1) = \Phi_u(1)D$ and further $\Sigma_u = DD'$, where $\Phi_\epsilon(1)$ is the long-run impact matrix of the underlying structural errors. The identifying restriction on $\Phi_\epsilon(1)$ is that only the technology shock has a permanent effect on labor productivity. This restriction implies that in our bivariate system the long-run impact matrix is triangular,

$$\Phi_\epsilon(1) = \begin{bmatrix} \Phi_{11} & 0 \\ \Phi_{21} & \Phi_{22} \end{bmatrix},$$

and it is assumed that $\Phi_{11} > 0$. Using $\Phi_\epsilon(1)\Phi'_\epsilon(1) = \Phi_u(1)\Sigma_u\Phi'_u(1)$ we can obtain $\Phi_\epsilon(1)$ from the Cholesky decomposition of $\Phi_u(1)\Sigma_u\Phi'_u(1)$. The contemporaneous impact matrix can be recovered from $D = [\Phi_u(1)]^{-1}\Phi_\epsilon(1)$. Correspondingly, the estimated versions are

$$\begin{aligned}\hat{\Phi}_\epsilon(1) &= \text{chol}[\hat{\Phi}_u(1)\widehat{\Sigma}_u\hat{\Phi}'_u(1)],\\ \hat{D} &= [\hat{\Phi}_u(1)]^{-1}\hat{\Phi}_\epsilon(1).\end{aligned}$$

Only the first column of $\hat{D}$ is identified and is our estimate of the first-period impact of the technology shock.

Next, we comment on the estimation techniques. First, note that for each representation there are several possible estimation methods. We chose algorithms that are both popular in the literature and known to work well in general. Of course, it is possible that there are algorithms that work slightly better for one of the representations in the current setting. However, the aim of this study is primarily to quantify whether the inclusion of the moving-average term alone leads to important gains in terms of more precise estimates of the structural parameters.

Vector Autoregressive Models: VARs are well known, so we comment only on a few issues. Fernández-Villaverde et al. (2005) show that for the CKM-specification there exists an infinite VAR representation. We verified that the same is true for the benchmark KP-specification. As in the previous Monte Carlo studies, the VAR lag length is set at four. However, for different sets of parameter values a VAR with different lags may yield slightly better results. We have chosen to stick to the VAR(4) because we want to facilitate comparison with the results of Christiano et al. (2006) and because there was no lag order that performed uniformly better for all DGPs.

State Space Models: There are many ways to estimate a state space model, e.g., the Kalman-based maximum likelihood methods and subspace identification methods such as N4SID of Van Overschee and DeMoor (1994) or the CCA method of Larimore (1983). An obvious candidate is maximum likelihood. Therefore, we included a prediction error method that is implemented with the PEM routine in the MATLAB system identification toolbox. However, it is well-known that maximum likelihood methods can face numerical problems that are due to the dependence on starting values, nonlinear optimization or local maxima. Indeed, these problems also apply to our setting. Therefore, we also use the CCA subspace algorithm that is asymptotically equivalent to maximum likelihood and was previously found to be remarkably accurate in small samples. As argued in Bauer (2005), CCA might be the best algorithm for econometric applications. The idea of subspace methods is that the state, $\mathbf{x}_t$, summarizes all information of the past that can be used for mean square prediction. Thus, the center of attention is the state that is estimated in a first step. In a second step the coefficient matrices are estimated by OLS. The different subspace algorithms use the structure of the state space representation in various ways. See Bauer (2005) for a more general introduction to subspace methods and the appendix for a detailed description of the algorithm that is employed in this paper.

While implementing the algorithm, we chose the correct dimension of the state vector, $n = 2$.[5] To calculate the long-run effect of the prediction errors, it is necessary to solve the state space equations $\mathbf{x}_{t+1} = A\mathbf{x}_t + Ku_t$, $\mathbf{y}_t = C\mathbf{x}_t + u_t$, where the deterministic component is omitted. The lag polynomial of the infinite moving-average representation is given by

$$\Phi_u(L) = I + \sum_{j=0}^{\infty} CA^j L^{j+1} K = I + LC(I - LA)^{-1}K.$$

An estimate of the long-run impact matrix $\Phi_u(1)$ can be obtained from the estimated system matrices, $\hat{\Phi}_u(1) = I + \hat{C}(I - \hat{A})^{-1}\hat{K}$. Henceforth, the estimation of the contemporaneous impact matrix is entirely analogous to long-run identification in a standard VAR setting. That is, we recover $\Phi_\epsilon(1)$ by a Cholesky decomposition and then obtain an estimate of D.

Vector Autoregressive Moving-Average Models: The VARMA representation given in (2.3) implies that we can represent $\mathbf{y}_t$ in terms of the innovations as

$$\mathbf{y}_t = (I - A_1 L)^{-1}(I + M_1 L)u_t = A(L)^{-1}M(L)u_t,$$

where $A(L)$ and $M(L)$ are the autoregressive polynomial and the moving-average polynomial, respectively, and the intercept term has been omitted. The long-run impact of u_t is given by $\Phi_u(1) = A(1)^{-1}M(1)$ and D can be recovered as before. The representation in (2.3) is however not the most useful representation in practice. It is more useful to choose a specific representation which guarantees that all parameters are identified and the number of estimated parameters is minimal. For an introduction to the identification problem in VARMA models see Lütkepohl (2005). Here we employ a final moving-average (FMA) representation that can be derived analogously to the final equation form (see Dufour and Pelletier, 2004). In our case, this results in a VARMA $(2, 1)$ representation in final moving-average form (see appendix).[6]

As in the case of state space models there are many different estimation methods for VARMA models. Examples are the methods developed by Hannan and Rissanen (1982), Koreisha and Pukkila (1990), Mauricio (1995) or Kapetanios (2003). We report results for a

[5] There are two auxiliary parameters in the subspace algorithm, f, $\mathfrak{p}$, which determine the row and column dimension of a Hankel matrix which is estimated in an intermediate step (see Bauer (2005) and the appendix). They have been set to $f = \mathfrak{p} = 8$. These parameters are of no importance asymptotically as long as they increase at certain rates with the sample size. In the literature it has been suggested to set $f = \mathfrak{p} = 2\hat{p}$ where $\hat{p}$ is the order of the chosen autoregressive approximation (Bauer, 2005).

[6] We experimented with other identified representations such as the final equation representation or the Echelon representation. However, the final moving-average representation always yielded the best results.

simple two-stage least squares method as in Hannan and Rissanen (1982), an iterative least squares estimation algorithm proposed by Kapetanios (2003) and the three-stage procedure developed by Hannan and Kavalieris (1984*b*). The two-stage least squares method starts with an initial "long" autoregression in order to estimate the unobserved residuals. The estimated residuals are then plugged into equation (2.3) and a (generalized) least squares regression is performed. The iterative least squares procedure takes these estimates as initial parameters and uses them to computes new residuals which are again used in a second least squares regression in order to update the parameter estimates. The updated parameter estimates are then used to update the residual estimates and so on, until convergence. The last algorithm is regression-based and is the first step of a Gauss-Newton procedure for the maximization of the likelihood, conditional on initial values. First, a high-order VAR is fitted to get initial estimates of the innovations. In the second stage these estimates are used to estimate the autoregressive and moving-average parameters by least squares. In the third stage the estimated coefficients are used to form new residuals and the coefficient estimates from the second stage are refined (see, e.g., Hannan and Kavalieris (1984*b*) or Hannan and Deistler (1988)). In the appendix we provide further details on the estimation algorithms. We use a VAR with lag length $n_T = 0.5\,\sqrt{T}$ for the initial long autoregression.[7]

2.4.2 Results of the Monte Carlo Study

Table 2.2 summarizes the results of the Monte Carlo simulation study. We tabulate Monte Carlo means and standard deviations of the estimates of the contemporaneous impact of a technology shock on productivity and hours worked for the various estimators. We also tabulate the MSE of the different estimators relative to the MSE of the estimator resulting from the benchmark SVAR. For the VARMA algorithms the estimation method is indicated in parenthesis, where $2SLS$ refers to the two-stage least squares method and ILS and $3SLS$ refer to the iterative least squares algorithm and the Hannan-Kavalieris method, respectively. Figures 2.1 and 2.2 depict the average estimated impulse responses of hours worked, together with the true impulse responses for the SVAR and the CCA subspace algorithm. In the figures, the bands around the mean lines correspond to the 0.025% and 0.975% quantiles of the estimated impulse responses at each point of time.

Our SVAR results confirm the findings of both Christiano et al. (2006) and Chari et al. (2005). While the SVAR is unbiased for the KP-specification (first row in table 2.2), the same is not true for the CKM-specification (fourth row in table 2.2). The associated pictures for both parameterizations show that the 95% bands around the mean impulse responses

[7] We also tried full information maximum likelihood maximization as, for example, in Mauricio (1995). However, this procedure proved to be highly unstable and was therefore not considered to be a practical alternative. One likely reason is that the roots of the AR and the MA polynomials are all close to the unit circle.

comprise a large region ranging from negative values to very high positive values. Also, for the different variations of the benchmark model we find that the SVAR is often biased and/or displays high variability. As can be seen from row 2, 3, 5 and 6 in table 2.2, both the biases and standard deviations are larger for the models with higher Frisch elasticities of labor supply (lower σ), as in the model this decreases the proportion of the variation in hours worked that is due to the technology shock. From row 7 and 8 it is clear that reducing the relative importance of the tax shock by lowering σ_l by 1/2 and 1/3 reduces the bias and the standard deviations.

The algorithms based on the state space representation perform quite differently. The PEM routine performs uniformly worse for all sets of parameter values. Although, in contrast to the SVAR, the state space model nests the DSGE model, the small sample performance of this estimation algorithm is much poorer. The low accuracy of the PEM routine can be attributed to the near non-stationarity and non-invertibility of the series that cause difficulties for the optimization procedure. The results of the PEM routine illustrate that using a formally exact representation of the DGP does not automatically lead to more precise estimates because the associated estimation algorithm may be numerically unreliable or not robust to the near violation of the underlying assumptions. For the CCA subspace algorithm, however, we find that the associated MSE of the estimated first-period impulse response is almost uniformly lower for both series and across different specifications. Only in two cases does the MSE of the CCA-based estimates exceed the MSE of the SVAR, and only by a very small amount. In particular, the first-period impact on hours worked is estimated more precisely up to a relative reduction to 85% in terms of MSE for the KP-specification. Figure 2.1 shows that the 95% interval is narrower for the estimated state space model, but still rather wide. In almost all cases the bias is at least slightly reduced. Although the response of hours worked is usually estimated more precisely, the performances of the subspace algorithm and the SVAR seem to be related: in cases where the SVAR does poorly, the state space model does so too. The advantage of the CCA algorithm over the SVAR-based least squares algorithm should be due to the use of the more general state space representation.[8]

The results for the different algorithms based on the VARMA representation are either similar to or worse than those for the VAR approximation. Generally, the less sophisticated methods give better results than the more complex estimation algorithms. Improvements in terms of bias are compromised by increases in variance and a higher MSE. The reason for this pattern is that we face an ill-conditioned problem that cannot be remedied by rescaling the data because of the presence of eigenvalues close to the unit circle of both the autoregressive and moving-average parts. Furthermore, the roots of the moving-average part and the autoregressive part imply that the model is close to being not identified. The iterative least

[8] It is also worth mentioning that since the CCA algorithm is based on OLS regressions, it is computationally not more intensive than a VAR. The same is not true for the PEM routine and most VARMA estimation algorithms.

squares method and the Hannan-Kavalieris method face consequently more problems than the simple two-stage least squares method. In comparison to the SVAR model, the structural estimates obtained from the VARMA algorithms perform relatively well in estimating the impact on hours worked, but worse in estimating the response of productivity to a technology shock. While the VARMA model fully nests the underlying DGP, this representation is not very efficient in our context.

A problem common to all algorithms is that the stationarity requirement is nearly violated for the DGPs at hand. As we have seen in section 2.4.1, the stationarity assumption lies at the heart of the long-run identification scheme. However, as the eigenvalues in table 2.1 indicate, this assumption is nearly violated for the benchmark models. This problem is independent of the chosen representation and, therefore, does not vanish even when we control for omitted moving-average terms. Apart from problems specific to the algorithms, this common problem may explain the relatively weak performance of all algorithms - a problem that could be overcome in larger samples.[9]

2.5 Conclusions

There has been some debate whether long-run identified SVARs can in practice discriminate between competing DSGE models and whether their sampling properties are good enough to justify their widespread use. Several Monte Carlo studies indicate that SVARs based on long-run restrictions are often biased and usually imprecise. Some authors have suggested that SVARs do poorly because they are only approximate representations of the underlying DGPs. Therefore, we replicate the simulation experiments of Chari et al. (2005) and Christiano et al. (2006) and apply more general models to their simulated data. In particular, we use algorithms based on VARMA and state space representations of the data and compare the resulting estimates of the underlying structural model. For our simulations, we found that one can do better by taking the full structure of the DGP into account. While our VARMA-based estimation algorithms and the prediction error algorithm for state space models were not found to do significantly better and often even worse, the CCA subspace algorithm seems to consistently outperform the SVAR. However, the estimates display high variability and are often biased, regardless of the reduced form model used. Furthermore, the performances of the different estimators are strongly correlated. This finding suggests that long-run identified SVARs do not fail because they are simple finite-order approximations. Instead, we find that the simulated processes are nearly violating the most basic assumptions on which long-run identification schemes are based. Given these properties of the data series, the poor performance seems almost entirely a small sample problem in this type of simulation studies.

[9] Our estimation results are in line with McGrattan's (2006) findings. However, McGrattan (2006) stresses the need to impose more theoretical restrictions to obtain informative statistics from empirical models.

Appendix

2.A Final MA Equation Form

Consider a standard representation for a stationary and invertible VARMA process

$$A(L)\mathbf{y}_t \quad = \quad M(L)u_t.$$

Recall that $M^{-1}(L) = M^*(L)/|M(L)|$, where $M^*(L)$ denotes the adjoint of $M(L)$ and $|M(L)|$ its determinant. We can multiply the above equation with $M^*(L)$ to get

$$M^*(L)A(L)\mathbf{y}_t \quad = \quad |M(L)|u_t.$$

This representation therefore places restrictions on the moving-average polynomial which is required to be a scalar operator, $|M(L)|$. Dufour and Pelletier (2004) show that this restriction leads to an identified representation. More specifically, consider the VARMA(1,1) representation in (2.3). Since the moving-average part is not of full rank we can write the system as

$$\begin{bmatrix} 1 - a_{11}L & -a_{12}L \\ -a_{21}L & 1 - a_{22}L \end{bmatrix} \mathbf{y}_t = \begin{bmatrix} 1 + m_{11}L & \alpha m_{11}L \\ m_{21}L & 1 + \alpha m_{21}L \end{bmatrix} u_t,$$

where α is some constant not equal to zero.
Clearly, $\det(M(L)) = 1 + (m_{11} + \alpha m_{21})L$ and we can write

$$\begin{bmatrix} 1 + \alpha m_{21}L & -\alpha m_{11}L \\ -m_{21}L & 1 + \alpha m_{11}L \end{bmatrix} \begin{bmatrix} 1 - a_{11}L & -a_{12}L \\ -a_{21}L & 1 - a_{22}L \end{bmatrix} \mathbf{y}_t = [1 + (m_{11} + \alpha m_{21})L]u_t.$$

Because of the reduced rank we end up with a VARMA $(2, 1)$. Note that the moving-average part is indeed restricted to be a scalar operator.

2.B Statistical Representations

This section elaborates on the derivation of the infinite VAR, VARMA and state space representations that result from our DSGE model in order to get an insight into the relationship between the economic model and the implied time series properties.

Consider again the law of motion of the logs

$$\begin{aligned}
\log k_{t+1} &= \phi_1 + \phi_{11} \log k_t - \phi_{11} \log x_t + \phi_{12}\tau_t, \\
\log y_t - \log L_t &= \phi_2 + \phi_{21} \log k_t - \phi_{21} \log x_t + \phi_{22}\tau_t, \\
\log L_t &= \phi_3 + \phi_{31} \log k_t - \phi_{31} \log x_t + \phi_{32}\tau_t,
\end{aligned}$$

and the exogenous states

$$\begin{aligned}
\log x_{t+1} &= \mu + \sigma_x \epsilon_{x,t+1}, \\
\tau_{t+1} &= (1-\rho)\bar{\tau}_l + \rho\tau_t + \sigma_l \epsilon_{l,t+1}.
\end{aligned}$$

From these equations the state space representation can be derived as follows. First write down the law of motion of labor productivity in differences:

$$\Delta \log(Y_t/L_t) = \log x_t + \phi_{21}\Delta \log k_t - \phi_{21}\Delta \log x_t + \phi_{22}\Delta\tau_t.$$

Thus the observed series can be expressed as

$$\begin{aligned}
\Delta \log(Y_t/L_t) &= \phi_{21} \log k_t - \phi_{21} \log k_{t-1} + (1-\phi_{21}) \log x_t \\
&\quad +\phi_{21} \log x_{t-1} + \phi_{22}\tau_t - \phi_{22}\tau_{t-1}, \\
\log L_t &= \phi_3 + \phi_{31} \log k_t - \phi_{31} \log x_t + \phi_{32}\tau_t.
\end{aligned}$$

Next, rewrite the law of motion for capital as

$$\log k_{t-1} = -\phi_{11}^{-1}\phi_1 + \phi_{11}^{-1} \log k_t + \log x_{t-1} - \phi_{11}^{-1}\phi_{12}\tau_{t-1},$$

in order to substitute for capital at time $t-1$:

$$\begin{aligned}
\Delta \log(Y_t/L_t) &= \phi_{21}\phi_{11}^{-1}\phi_1 + \phi_{21}(1-\phi_{11}^{-1}) \log k_t \\
&\quad +(1-\phi_{21}) \log x_t + \phi_{22}\tau_t + (\phi_{21}\phi_{11}^{-1}\phi_{12} - \phi_{22})\tau_{t-1}.
\end{aligned}$$

Using the laws of motion for the stochastic shock processes, substitute the current exogenous shocks to get

$$\begin{aligned}
\Delta \log(Y_t/L_t) &= \left[\phi_{21}\phi_{11}^{-1}\phi_1 + (1-\phi_{21})\mu + \phi_{22}(1-\rho)\bar{\tau}_l\right] + \phi_{21}(1-\phi_{11}^{-1})\log k_t \\
&\quad +(\phi_{21}\phi_{11}^{-1}\phi_{12} - (1-\rho)\phi_{22})\tau_{t-1} + (1-\phi_{21})\sigma_x\epsilon_{x,t} + \phi_{22}\sigma_l\epsilon_{l,t}, \\
\log L_t &= [\phi_3 - \phi_{31}\mu + \phi_{32}(1-\rho)\bar{\tau}_l] + \phi_{31}\log k_t + \phi_{32}\rho\tau_{t-1} \\
&\quad -\phi_{31}\sigma_x\epsilon_{x,t} + \phi_{32}\sigma_l\epsilon_{l,t}.
\end{aligned}$$

Next, consider the law of motion for capital and express future capital in terms of the current states as

$$\begin{aligned}
\log k_{t+1} &= [\phi_1 - \phi_{11}\mu + \phi_{12}(1-\rho)\bar{\tau}_l] + \phi_{11}\log k_t + \phi_{12}\rho\tau_{t-1} \\
&\quad -\phi_{11}\sigma_x\epsilon_{x,t} + \phi_{12}\sigma_l\epsilon_{l,t}.
\end{aligned}$$

Collecting the above equations, the system can be written in state space form according to Fernández-Villaverde et al. (2005). The state transition equation is

$$\begin{bmatrix} \log k_{t+1} \\ \tau_t \end{bmatrix} = K_1 + A \begin{bmatrix} \log k_t \\ \tau_{t-1} \end{bmatrix} + B \begin{bmatrix} \epsilon_{x,t} \\ \epsilon_{lt} \end{bmatrix},$$

where the system matrices are given by

$$K_1 = \begin{bmatrix} \phi_1 - \phi_{11}\mu + \phi_{12}(1-\rho)\bar{\tau}_l \\ (1-\rho)\bar{\tau} \end{bmatrix},$$

$$A = \begin{bmatrix} \phi_{11} & \phi_{12}\rho \\ 0 & \rho \end{bmatrix},$$

and

$$B = \begin{bmatrix} -\phi_{11}\sigma_x & \phi_{12}\sigma_l \\ 0 & \sigma_l \end{bmatrix}.$$

The observation equation is

$$\begin{bmatrix} \Delta \log(Y_t/L_t) \\ \log L_t \end{bmatrix} = K_2 + C \begin{bmatrix} \log k_t \\ \tau_{t-1} \end{bmatrix} + D \begin{bmatrix} \epsilon_{x,t} \\ \epsilon_{lt} \end{bmatrix},$$

with system matrices

$$K_2 = \begin{bmatrix} \phi_{21}\phi_{11}^{-1}\phi_1 + (1-\phi_{21})\mu + \phi_{22}(1-\rho)\bar{\tau}_l \\ \phi_3 - \phi_{31}\mu + \phi_{32}(1-\rho)\bar{\tau}_l \end{bmatrix},$$

$$C = \left[\begin{array}{cc} \phi_{21}(1-\phi_{11}^{-1}) & \phi_{21}\phi_{11}^{-1}\phi_{12} - (1-\rho)\phi_{22} \\ \phi_{31} & \phi_{32}\rho \end{array} \right],$$

and

$$D = \left[\begin{array}{cc} (1-\phi_{21})\sigma_x & \phi_{22}\sigma_l \\ -\phi_{31}\sigma_x & \phi_{32}\sigma_l \end{array} \right].$$

This representation permits us to derive the infinite VAR and VARMA representation in compact form.

Let $\mathbf{y}_t$ denote the vector of observables, $\mathbf{x}_t$ the vector of states, and ϵ the white noise shocks. Then we have as above

$$\begin{aligned} \mathbf{x}_{t+1} &= K_1 + A\mathbf{x}_t + B\epsilon_t, \\ \mathbf{y}_t &= K_2 + C\mathbf{x}_t + D\epsilon_t. \end{aligned}$$

If D is invertible, it is possible to use

$$\epsilon_t = D^{-1}(\mathbf{y}_t - K_2 - C\mathbf{x}_t)$$

in the transition equation to obtain

$$\begin{aligned} \mathbf{x}_{t+1} &= K_1 + A\mathbf{x}_t + BD^{-1}(\mathbf{y}_t - K_2 - C\mathbf{x}_t), \\ (I - (A - BD^{-1}C)L)\mathbf{x}_{t+1} &= [K_1 - BD^{-1}K_2] + BD^{-1}\mathbf{y}_t. \end{aligned}$$

If the eigenvalues of $(A - BD^{-1}C)$ are strictly less than one in modulus we can solve for $\mathbf{x}_{t+1}$:

$$\mathbf{x}_{t+1} = \left(I - (A - BD^{-1}C)L\right)^{-1}\left([K_1 - BD^{-1}K_2] + BD^{-1}\mathbf{y}_t\right).$$

Using this relation in the observation equation yields the infinite VAR representation for $\mathbf{y}_t$:

$$\mathbf{y}_t = K_2 + C\left(I - (A - BD^{-1}C)L\right)^{-1}\left([K_1 - BD^{-1}K_2] + BD^{-1}\mathbf{y}_{t-1}\right) + D\epsilon_t,$$

$$\mathbf{y}_t = K_3 + C\left(I - (A - BD^{-1}C)L\right)^{-1} BD^{-1}\mathbf{y}_{t-1} + D\epsilon_t.$$

Note that the condition for the existence of an infinite VAR-representation is that $I - (A - BD^{-1}C)$ is invertible. If this condition does not hold, impulse responses from a VAR are unlikely to match up those from the model.

If C is invertible, it is possible to rewrite the state as

$$\mathbf{x}_t = C^{-1}\left(\mathbf{y}_t - K_2 - D\epsilon_t\right)$$

and use it in the transition equation:

$$C^{-1}\left(\mathbf{y}_{t+1} - K_2 - D\epsilon_{t+1}\right) = K_1 + AC^{-1}\left(\mathbf{y}_t - K_2 - D\epsilon_t\right) + B\epsilon_t,$$

$$\mathbf{y}_{t+1} - CAC^{-1}\mathbf{y}_t = CK_1 + K_2 - CAC^{-1}K_2 + (CB - CAC^{-1}D)\epsilon_t + D\epsilon_{t+1}.$$

Therefore, we obtain a VARMA(1,1) representation of $\mathbf{y}_t$:

$$\mathbf{y}_t = K_4 + CAC^{-1}\mathbf{y}_{t-1} + \left(I + (CBD^{-1} - CAC^{-1})L\right)u_t.$$

with $u_t \sim N(0, DD')$.

2.C Estimation Algorithms

2.C.1 Two-State Least Squares and Iterative Least Squares

These two methods are computationally very easy to implement. The iterative least squares method has been introduced by Kapetanios (2003).

We discuss the methods in the framework of a standard VARMA (p,q) representation

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

Usually additional restrictions need to be imposed on the coefficient matrices to ensure identification of the parameters.

Given that the moving-average polynomial is invertible, there exists an infinite VAR representation of the process, $y_t = \sum_{i=1}^{\infty} \Pi_i y_{t-i} + u_t$. In the first step of both algorithms, this representation is approximated by a “long” VAR to get an estimate of the residuals. More precisely, the following regression equation is used

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

where n_T is large and goes to infinity as the sample size grows. The estimated residuals are denoted by $\hat{u}_t^{(0)}$. Given these estimates, we might obtain estimates of the parameter matrices by performing a (restricted) regression in

$$y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t + M_1 \hat{u}_{t-1}^{(0)} + \ldots + M_q \hat{u}_{t-q}^{(0)}.$$

Denote the estimated coefficient matrices by $A_1^{(1)}, A_2^{(1)}, \ldots, A_p^{(1)}$ and $M_1^{(1)}, M_2^{(1)}, \ldots, M_q^{(1)}$. These estimates are the final estimates of the two-stage least squares method. These initial parameter estimates can be used to obtain a new estimate of the residuals. Denote by $\hat{U}^{(1)}$ the vector collecting the estimated new residuals. We can then use $\hat{U}^{(1)}$ again in the above equation to obtain new estimates of the coefficient matrices. Denote the vector of estimated residuals at the i^{th} iteration by $\hat{U}^{(i)}$. Kapetanios (2003) proposes to iterate least squares regressions until $||U^{(i-1)} - \hat{U}^{(i)}|| < c$, according to some pre-specified number c.

2.C.2 Hannan-Kavalieris Method

This method goes originally back to Durbin (1960) and has been introduced by Hannan and Kavalieris (1984*a*) for multivariate processes.[10] 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 and Pelletier (2004)). We discuss also this method in the framework of a standard VARMA (p, q) representation

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

To consider zero restrictions, we use the following notation. The vector of all parameters is denoted by $\beta = \text{vec}[A_1, \ldots, A_p, M_1, \ldots M_q]$. The vector of free parameters, γ, can be defined by introducing a restriction matrix R such that the vectors are related by $\beta = R\gamma$.

Given that the moving-average polynomial is invertible, there exists an infinite VAR representation of the process, $y_t = \sum_{i=1}^{\infty} \Pi_i y_{t-i} + u_t$. In the first step of the algorithm, this representation is approximated by a "long" VAR to get an estimate of the residuals. More precisely, the following regression equation is used

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

where n_T is large and goes to infinity as the sample size grows. Given an estimate of the residuals, $\hat{u}_t$, we might obtain starting values for future iterations by performing a (restricted) regression in

$$y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t + M_1 \hat{u}_{t-1} + \ldots + M_q \hat{u}_{t-q}.$$

Denote the estimated coefficient matrices by $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_p$ and $\tilde{M}_1, \tilde{M}_2, \ldots, \tilde{M}_q$. The first iteration of the conditional maximum likelihood algorithm can be expressed in a simple regression framework. One forms new residuals, ε_t, and new matrices, ξ_t, η_t and $\hat{X}_t$, according

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

to

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

for $t = 1, 2, \dots, T$, $Y_t = [y_t', \dots, y_{t-p+1}', \hat{u}_t', \dots, \hat{u}_{t-q+1}']'$ and $y_t = \varepsilon_t = \xi_t = \eta_t = 0$ and $\hat{X}_t = 0$ for $t \leq 0$. The final estimate is

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

where $\widehat{\Sigma} = T^{-1} \sum \varepsilon_t \varepsilon_t'$, $m = \max\{p, q\}$. This procedure is asymptotically efficient under certain conditions (Lütkepohl (2005)).

2.C.3 Subspace Algorithms

Subspace algorithms rely on the state space representation of a linear system. The CCA algorithm is originally due to Larimore (1983). The basic idea behind subspace algorithms lies in the fact that if we knew the unobserved state, x_t, we could estimate the *system matrices*, A, K, C, by linear regressions as can be seen from the basic equations

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

Given the state and the observations, $\hat{C}$ and $\hat{u}_t$ could be obtained by a regression of y_t on x_t and $\hat{A}$ and $\hat{K}$ could be obtained by a regression of x_{t+1} on x_t and $\hat{u}_t$. Therefore, the problem is to obtain 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 unobserved residuals in a first step which 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

$$\begin{aligned} x_t &= (A - KC)^{\mathfrak{p}} x_{t-\mathfrak{p}} + \sum_{i=0}^{\mathfrak{p}-1} (A - KC)^i K y_{t-i-1}, \\ &= (A - KC)^{\mathfrak{p}} x_{t-\mathfrak{p}} + \mathcal{K}_{\mathfrak{p}} Y_{t,\mathfrak{p}}^-, \end{aligned} \tag{2.4}$$

where $\mathcal{K}_{\mathfrak{p}} = [K, (A - KC)K, \ldots, (A - KC)^{\mathfrak{p}-1}K]$ 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

$$y_{t+j} = CA^j x_t + \sum_{i=0}^{j-1} CA^i K u_{t+j-i-1} + u_{t+j}, \tag{2.5}$$

for $j = 1, 2, \ldots$. 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 this sense all relevant 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 (2.5) for all observations contained in $Y_{t,f}^+$ simultaneously. Combine these equations with (2.4) in order to obtain

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

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 $(A - KC)$ is less than one in absolute value, we have $(A - KC)^{\mathfrak{p}} \approx 0$ for large $\mathfrak{p}$. This condition is satisfied for stationary and invertible processes. This reasoning motivates an approximation of the above equation by

$$Y_{t,f}^+ = \beta Y_{t,\mathfrak{p}}^- + N_{t,f}^+, \tag{2.6}$$

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 that is decomposed into $\mathcal{O}_f$ and $\mathcal{K}_{\mathfrak{p}}$. The identification problem is solved implicitly during this step.

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 (2.6) 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

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

   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

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

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 prior to estimation, namely the dimension of the state and the integers f and $\mathfrak{p}$. See the text for the employed values. For the asymptotic consequences of various choices see Bauer (2005).

2.D Tables and Figures

Table 2.1: Benchmark Calibrations and Time Series Properties

Parameters	Common	CKM Benchmark	KP Benchmark
α	0.33		
β	$0.98^{1/4}$		
σ	1		
δ	$1-(1-0.6)^{1/4}$		
ψ	2.5		
γ	$1.01^{1/4}-1$		
μ	0.00516		
$\bar{L}$	1		
$\bar{\tau}_l$	0.243		
τ_x	0.3		
ρ		0.94	0.993
σ_τ		0.008	0.0066
σ_x		0.00568	0.011738
Selected time series properties			
eig(A_1)		$0.9573,\ 0.9400$	$0.9573,\ 0.9930$
eig(M_1)		$-0.9557,\ 0$	$-0.9505,\ 0$
eig(K)		$-1.7779 \pm 0.51i$	$-2.0298 \pm 0.35i$

Parameter values of the CKM and KP benchmark calibrations. The last three rows display some properties of the implied VARMA and state space representation. eig(A_1) and eig(M_1) denote the eigenvalues of the autoregressive and the moving-average matrix, respectively. eig(K) denotes the eigenvalues of the matrix K in the state space model in innovations form.

Table 2.2: Simulation Results

		VAR (4)			PEM (2)			SS (2,8,8)			VARMA (2SLS)			VARMA (ILK)			VARMA (3SLS)		
Variable	True Value	Mean	Std.	MSE	Mean	Std.	MSE	Mean	Std.	MSE	Mean	Std.	MSE	Mean	Std.	MSE	Mean	Std.	MSE
KP Benchmark																			
Prod.	0.69	0.55	0.19	1	0.53	0.23	1.44	0.57	0.18	0.86	0.54	0.18	1.02	0.54	0.19	1.06	0.53	0.21	1.23
Hours	0.28	0.31	0.43	1	0.32	0.50	1.33	0.31	0.40	0.85	0.33	0.42	0.92	0.32	0.42	0.94	0.36	0.47	1.22
KP, $\sigma = 0$ (Indivisible Labor)																			
Prod.	0.65	0.48	0.23	1	0.45	0.28	1.49	0.50	0.22	0.92	0.47	0.22	1.02	0.47	0.22	1.02	0.47	0.23	1.09
Hours	0.43	0.56	0.56	1	0.58	0.67	1.44	0.52	0.54	0.92	0.57	0.54	0.93	0.56	0.55	0.98	0.62	0.63	1.30
KP, $\sigma = 6$ (Frisch elasticity=0.63)																			
Prod.	0.75	0.61	0.16	1	0.57	0.24	2.03	0.63	0.14	0.80	0.59	0.16	1.10	0.59	0.16	1.16	0.59	0.17	1.26
Hours	0.11	0.10	0.19	1	0.10	0.24	1.52	0.10	0.18	0.89	0.10	0.19	0.98	0.10	0.19	1.00	0.11	0.23	1.37
CKM Benchmark																			
Prod.	0.34	0.10	0.17	1	0.13	0.20	1.03	0.11	0.18	1.00	0.09	0.16	1.07	0.09	0.17	1.08	0.10	0.17	1.04
Hours	0.14	0.65	0.39	1	0.61	0.53	1.21	0.62	0.40	0.95	0.67	0.37	1.02	0.66	0.40	1.03	0.70	0.41	1.18
CKM, $\sigma = 0$ (Indivisible Labor)																			
Prod.	0.31	-0.12	0.21	1	-0.07	0.30	1.02	-0.12	0.23	1.05	-0.15	0.19	1.09	-0.13	0.21	1.06	-0.12	0.22	1.04
Hours	0.21	1.26	0.49	1	1.16	0.84	1.20	1.24	0.54	1.01	1.31	0.45	1.06	1.28	0.50	1.03	1.33	0.55	1.17
CKM, $\sigma = 6$ (Frisch elasticity=0.63)																			
Prod.	0.36	0.30	0.08	1	0.29	0.11	1.67	0.31	0.08	0.95	0.29	0.08	1.04	0.29	0.09	1.21	0.30	0.09	1.24
Hours	0.05	0.12	0.17	1	0.12	0.20	1.35	0.11	0.17	0.92	0.13	0.17	0.97	0.12	0.17	1.02	0.14	0.18	1.11
CKM, $\sigma_l/2$																			
Prod.	0.34	0.25	0.10	1	0.25	0.12	1.45	0.26	0.10	0.93	0.25	0.09	1.06	0.25	0.10	1.07	0.25	0.10	1.17
Hours	0.14	0.26	0.22	1	0.26	0.26	1.35	0.24	0.21	0.88	0.27	0.21	0.97	0.26	0.22	1.01	0.28	0.23	1.19
CKM, $\sigma_l/3$																			
Prod.	0.34	0.28	0.07	1	0.28	0.09	1.63	0.29	0.07	0.89	0.28	0.07	1.08	0.28	0.07	1.13	0.28	0.07	1.11
Hours	0.14	0.18	0.15	1	0.18	0.18	1.44	0.17	0.14	0.87	0.19	0.14	0.96	0.18	0.15	1.01	0.20	0.16	1.14

Percent contemporaneous impact on productivity and hours of one standard deviation shock to technology.

The entries are Monte Carlo means and standard deviations. MSEs are relative to the MSE of the SVAR.

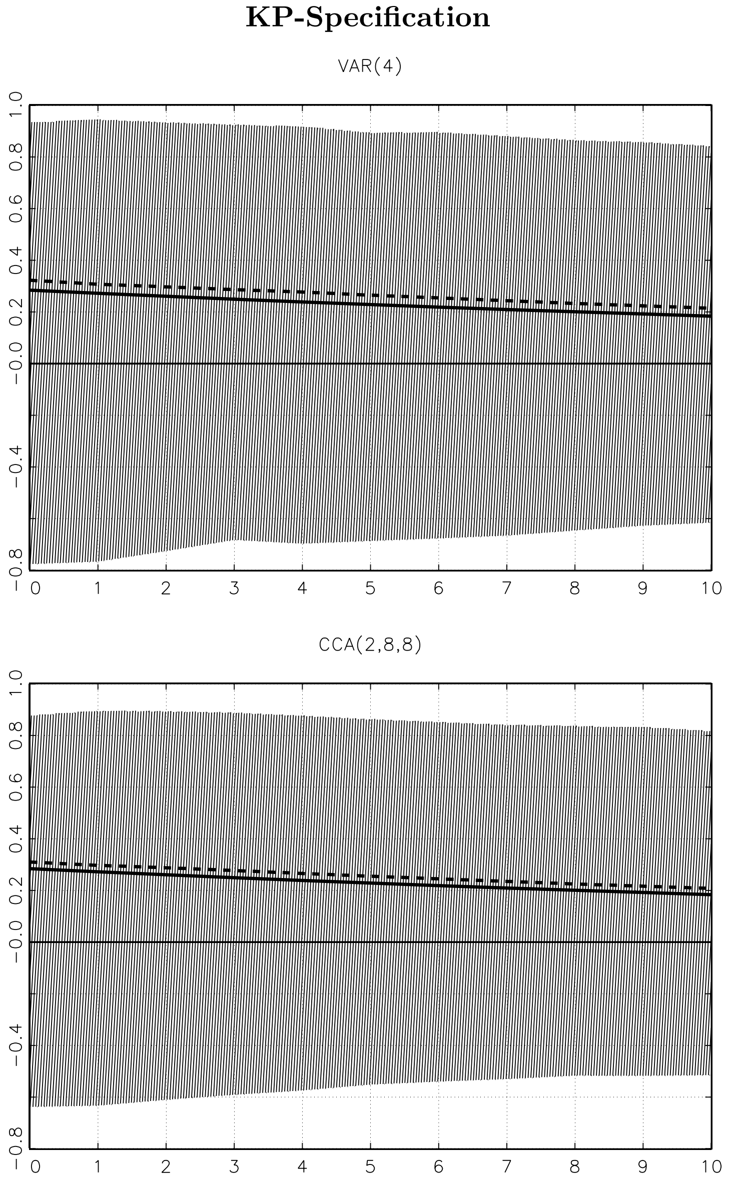

Figure 2.1: Mean impulse response (- -), true impulse response (–) and 95% intervals of hours worked to one standard deviation shock to technology for the VAR and the CCA subspace algorithm.

CKM-Specification

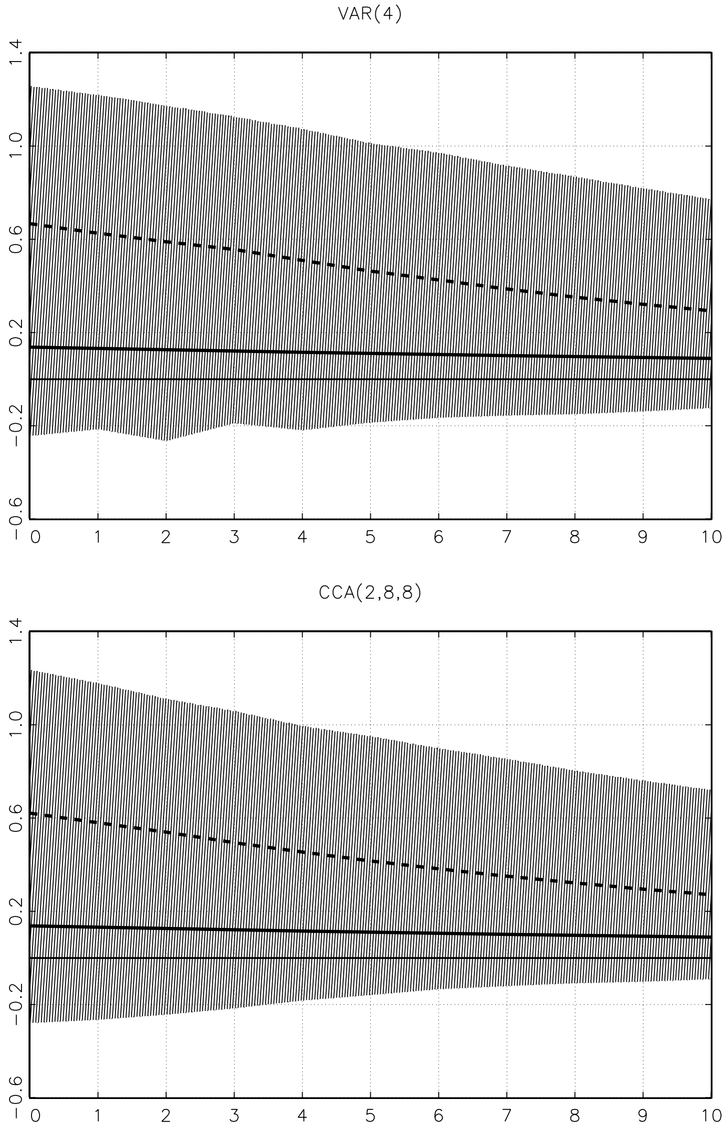

Figure 2.2: Mean impulse response (- -), true impulse response (–) and 95% intervals of hours worked to one standard deviation shock to technology for the VAR and the CCA subspace algorithm.

Bibliography

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

Blanchard, O. J. and Quah, D. (1989), ‘The dynamic effects of aggregate demand and supply disturbances’, *American Economic Review* **79**(4), 655–673.

Chari, V. V., Kehoe, P. J. and McGrattan, E. R. (2005), ‘A critique of structural VARs using real business cycle theory’. Federal Reserve Bank of Minneapolis, Working Paper 631, May draft.

Christiano, L. and Eichenbaum, M. (1992), Identification and the liquidity effect of a monetary policy shock, *in* A. Cukierman, Z. Hercowitz and L. Leiderman, eds, ‘Political Economy, Growth and Business Cycles’, MIT Press, Cambridge and London, pp. 335–370.

Christiano, L. J., Eichenbaum, M. and Vigfusson, R. (2006), ‘Assessing structural VARs’. Northwestern University and NBER, Federal Reserve Board of Governors, Working Paper, March draft.

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

Dufour, J.-M. and 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.

Erceg, C. J., Guerrieri, L. and Gust, C. (2005), ‘Can long-run restrictions identify technology shocks?’. International Financial Discussion Paper 792, March draft.

Fernández-Villaverde, J., Rubio-Ramírez, J. and Sargent, T. J. (2005), ‘A,B,C’s (and D)’s for understanding VARs’. NBER Technical Working Paper 308, May draft.

Gali, J. (1999), 'Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations?', *American Economic Review* **89**(1), 249–271.

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

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

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

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

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

King, R. G., Plosser, C. I., Stock, J. H. and Watson, M. W. (1991), 'Stochastic trends and economic fluctuations', *American Economic Review* **81**(4), 819–840.

Klein, A., Mélard, G. and Spreij, P. (2004), 'On the resultant property of the Fisher information matrix of a vector ARMA process.'. Universiteit van Amsterdam, Discussion Paper 2004/13.

Koreisha, S. and 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 Modeling via Canonical Variate Analysis, *in* H. S. Rao and P. Dorato, eds, 'Proc. 1983 Amer. Control Conference 2'.

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.

McGrattan, E. R. (2006), 'Measurement with Minimal Theory'. Federal Reserve Bank of Minneapolis, Working Paper 643, July draft.

Prescott, E. C. (1986), 'Theory ahead of business cycle measurement', pp. 9–22. Quarterly Review, Federal Reserve Bank of Minneapolis, issue Fall.

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

Chapter 3

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

3.1 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 and 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 et al. (2005) show formally how DSGE models and VARMA processes are linked. Also Cooley and 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 and 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 and 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 and 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 and 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 and Pukkila (1990), a three-stage least squares procedure based on Hannan and Kavalieris (1984*a*) and the CCA subspace algorithm by Larimore (1983). The obtained results suggest that the algorithm of Hannan and 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.

3.2 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

$$A_0 y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + M_0 u_t + M_1 u_{t-1} + \ldots + M_q u_{t-q} \tag{3.1}$$

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 (3.1) 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

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

subtracted prior to estimation. We will also abstract from issues such as seasonality. As will be seen later, we consider models of the form (3.1) 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 (3.1) more compactly as

$$A(L)y_t \quad = \quad M(L)u_t,\ t \in \mathbb{Z}. \tag{3.2}$$

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

$$y_t \quad = \quad \sum_{i=0}^{\infty} \Phi_i u_{t-i} = \Phi(L)u_t, \tag{3.3}$$

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

$$y_t \quad = \quad \sum_{i=1}^{\infty} \Pi_i y_{t-i} + u_t, \tag{3.4}$$

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 (3.1) 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

$$D(L)[\bar{A}(L), \bar{M}(L)] = [A(L), M(L)] \tag{3.5}$$

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 (3.5) 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 and 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.

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

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, $\alpha_{ki}(L)$, $k \neq i$. Again, it can be shown that this representation leads to identified parameters (Hannan and 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 (3.1) can also be written as a state space model of the form

$$\begin{aligned} x_{t+1} &= A x_t + B u_t, \\ y_t &= C x_t + u_t, \end{aligned} \tag{3.6}$$

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 (3.6) can be shown to have a VARMA representation (see, e.g., Aoki, 1989; Hannan and 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 and 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 (3.6) 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 $E[x_tx_t'] = \mathcal{K}E[Y_t^-(Y_t^-)']\mathcal{K}' = \mathcal{K}\Gamma_\infty^-\mathcal{K}'$, where $\Gamma_\infty^- := E[Y_t^-(Y_t^-)']$. Given a state space system as in (3.6), there is also another representation which is called *backward innovation model*

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

where time "runs backwards" and A is as in (3.6) and N $(n \times K)$, M $(n \times K)$ are functions of (A, B, C), in particular $M = E[x_ty_{t-1}']$. 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 $E[z_tz_t'] = \mathcal{O}'(E[Y_t^+(Y_t^+)'])^{-1}\mathcal{O} = \mathcal{O}'(\Gamma_\infty^+)^{-1}\mathcal{O}$, 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 $E[x_tx_t'] = E[z_tz_t'] = diag(\sigma_1, \ldots, \sigma_n)$, 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

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

where $\gamma(j) := E[y_t y_{t-j}'],\ j = 1, 2, \dots$ 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 and 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.

3.3 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 (3.1) 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

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

$y_1, y_2, \ldots, y_T$. 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

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

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

$$\beta = R\gamma. \tag{3.7}$$

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 (3.4) 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

$$y_t = \sum_{i=1}^{n_T} \Pi_i y_{t-i} + u_t, \tag{3.8}$$

where n_T is the lag length that is required to increase with the sample size, T. In the second stage, the residuals from (3.8), $\hat{u}_t^{(0)}$, are plugged in (3.1). After rearranging (3.1), one gets

$$\begin{aligned} y_t &= (I_K - A_0)[y_t - \hat{u}_t^{(0)}] + A_1 y_{t-1} + \ldots + A_p y_{t-p} \\ &\quad + M_1 \hat{u}_{t-1}^{(0)} + \ldots + M_q \hat{u}_{t-q}^{(0)} + u_t, \end{aligned} \tag{3.9}$$

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

$$Y = [I_K - A_0, \mathbf{A}, \mathbf{M}]X^{(0)} + U, \tag{3.10}$$

where $Y := [y_{n_T+m+1}, \dots, y_T]$, $U := [u_{n_T+m+1}, \dots, u_T]$ is the matrix of regression errors, $X^{(0)} := [Y_{n_T+m}^{(0)}, \dots, Y_{T-1}^{(0)}]$ 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 (3.10) yields

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

and the 2SLS estimator is defined as

$$\tilde{\gamma} = [R'(X^{(0)}X^{(0)'} \otimes \widetilde{\Sigma}^{-1})R]^{-1}R'(X^{(0)} \otimes \widetilde{\Sigma}^{-1})\text{vec}(Y). \tag{3.11}$$

where $\widetilde{\Sigma}$ is the covariance matrix estimator based on the residuals $\hat{u}_t^{(0)}$. The corresponding estimated matrices are denoted by $\tilde{A}_0, \tilde{A}_1, \dots, \tilde{A}_p$ and $\tilde{M}_1, \tilde{M}_2 \dots, \tilde{M}_q$, respectively. Alternatively, one may also plug in the estimated current innovation $\hat{u}_t^{(0)}$ in (3.9), 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 and Pukkila, 1989).

For univariate and multivariate models different selection rules for the lag length of the initial autoregression have been proposed. For example, Hannan and Kavalieris (1984a) propose to select n_T by AIC or BIC, while Koreisha and 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 and Pukkila, 1990). Lütkepohl and 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]

[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

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 and 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 and 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

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

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 and Kavalieris (1984*b*) showed consistency and asymptotic normality of these estimators. Dufour and Pelletier (2004) extend these results to even more general conditions. The Monte Carlo evidence presented by Dufour and 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

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 and Deistler (1988), sections 6.5, 6.7, for an extensive discussion.

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 and Pukkila (1990a) proposed this procedure for univariate ARMA models and Kavalieris, Hannan and Salau (2003) proved efficiency of the GLS estimates in this case. See also Flores de Frutos and 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 and 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 (3.8). Koreisha and Pukkila (1990a) *assume* that the residuals obtained from (3.8) estimate the true residuals up to an uncorrelated error term, $u_t = \hat{u}_t^{(0)} + \epsilon_t$. If this expression is inserted in (3.1), one obtains

$$\begin{aligned}
A_0 y_t &= \sum_{j=1}^{p} A_j y_{t-j} + A_0(\hat{u}_t^{(0)} + \epsilon_t) + \sum_{j=1}^{q} M_j(\hat{u}_{t-j}^{(0)} + \epsilon_{t-j}), \\
y_t &= (I_K - A_0)(y_t - \hat{u}_t^{(0)}) + \sum_{j=1}^{p} A_j y_{t-j} + \hat{u}_t^{(0)} \\
&\quad + \sum_{j=1}^{q} M_j \hat{u}_{t-j}^{(0)} + A_0 \epsilon_t + \sum_{j=1}^{q} M_j \epsilon_{t-j}, \\
y_t - \hat{u}_t^{(0)} &= (I - A_0)(y_t - \hat{u}_t^{(0)}) + \sum_{j=1}^{p} A_j y_{t-j} \\
&\quad + \sum_{j=1}^{q} M_j \hat{u}_{t-j}^{(0)} + \zeta_t. \qquad (3.12)
\end{aligned}$$

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 (3.12) is not efficient. Koreisha and 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 (3.12) 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

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

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

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

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 $\zeta = (\zeta_{n_T+m+1}' \cdots \zeta_T')'$. Let $\Phi := E[\zeta\zeta']$ and denote its estimate by $\hat{\Phi}$. In the third stage, we re-estimate (3.12) by GLS using $\hat{\Phi}$:

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

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

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

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 (3.11) in

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

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 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.

[6] The evidence given by Koreisha and 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.

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 and 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

$$\begin{aligned} x_t &= (A - BC)^{\mathfrak{p}} x_{t-\mathfrak{p}} + \sum_{i=0}^{\mathfrak{p}-1} (A - BC)^i B y_{t-i-1}, \\ &= (A - BC)^{\mathfrak{p}} x_{t-\mathfrak{p}} + \mathcal{K}_{\mathfrak{p}} Y_{t,\mathfrak{p}}^{-}, \end{aligned} \tag{3.13}$$

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

$$y_{t+j} = CA^j x_t + \sum_{i=0}^{j-1} CA^i B u_{t+j-i-1} + u_{t+j}. \tag{3.14}$$

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 (3.14) for all observations contained in $Y_{t,f}^{+}$ simultaneously. Combine these equations with (3.13) in order to obtain

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

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

$$Y_{t,f}^{+} = \beta Y_{t,\mathfrak{p}}^{-} + N_{t,f}^{+} \tag{3.15}$$

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 (3.15) 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

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

   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.

3.4 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 3.4.3. 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.

3.4.1 Performance Measures

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:

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

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 $h = 1, \dots, q$. 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.

Impulse Response Analysis

Researchers might also be interested in the accuracy of the estimated impulse response function as in (3.3),

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

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.

3.4.2 Generated Systems

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 3.1 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.

Higher-Dimensional Systems

DGP III: I consider a three-dimensional system that was used extensively in the literature by, e.g., Koreisha and Pukkila (1989), Flores de Frutos and Serrano (2002) and others for illustrative purposes. Koreisha and 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 and Pukkila (1989) and has eigenvalues $\lambda^{ar} = 0.7$ and $\lambda^{ma}_1 = -0.6$ and $\lambda^{ma}_2 = 0.5$.

DGP IV: This process has been used in the simulation studies of Koreisha and 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 and Pukkila (1987). That is, $\alpha_{11,1} = 0.5$ and $m_{55,1} = -0.6$ with eigenvalues $\lambda^{ar}_1 = 0.5$, $\lambda^{ar}_2 = 0 \pm i0.57$, $\lambda^{ma}_1 = -0.6$ and $\lambda^{ma}_2 = -0.4 \pm i0.67$.

3.4.3 Results

The results are summarized in tables 3.2 to 3.5 and figures 3.1 to 3.8. The tables show the frequency of cases when the algorithms failed for different reasons. The figures plot the various MSE ratios discussed above.

Table 3.2 and table 3.3 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 3.4 and 3.5 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 3.4 and 3.5.

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.

3.5 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 and 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 and 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 and 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 and 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.

Appendix

3.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 $y_t = (y_{1t}, \dots, y_{Kt})'$ 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

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

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

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

[7] See also the book of Hannan and 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 $D_i = c_{n-i}I_K + \sum_{k=1}^{n-i} c_{n-i-k}CA^{k-1}B$. According to the Cayley - Hamilton theorem, the matrix polynomial in A vanishes, $A^n + c_1A^{n-1} + \ldots + c_nI_n = 0$ (Aoki, 1989). One obtains therefore the following VARMA representation

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

3.B Figures and Tables

Table 3.1: 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 3.2: 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 3.3: 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 3.4: 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 3.5: 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.

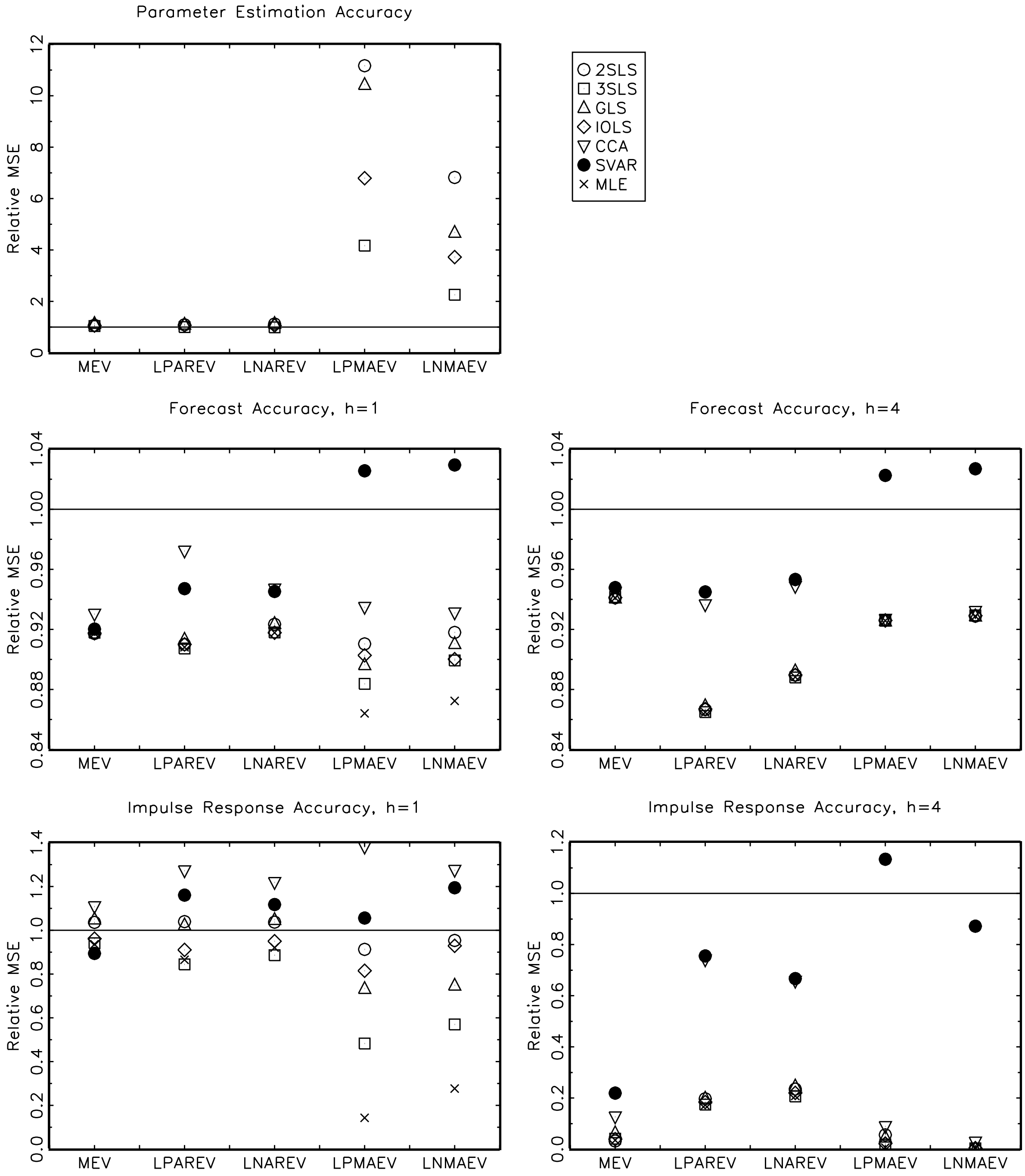

Figure 3.1: MSE ratios for DGP I with $T = 100$.

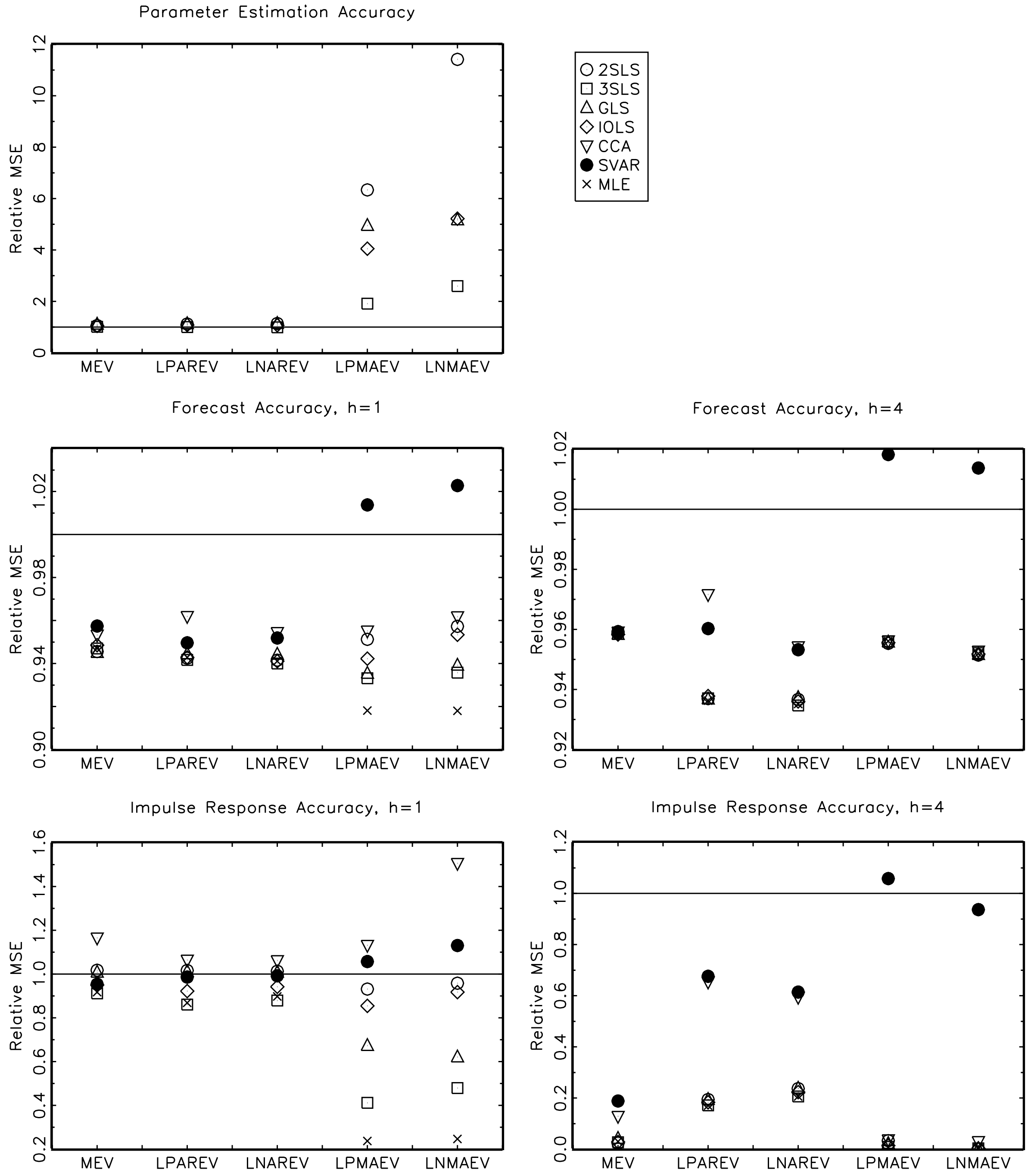

Figure 3.2: MSE ratios for DGP I with $T = 200$.

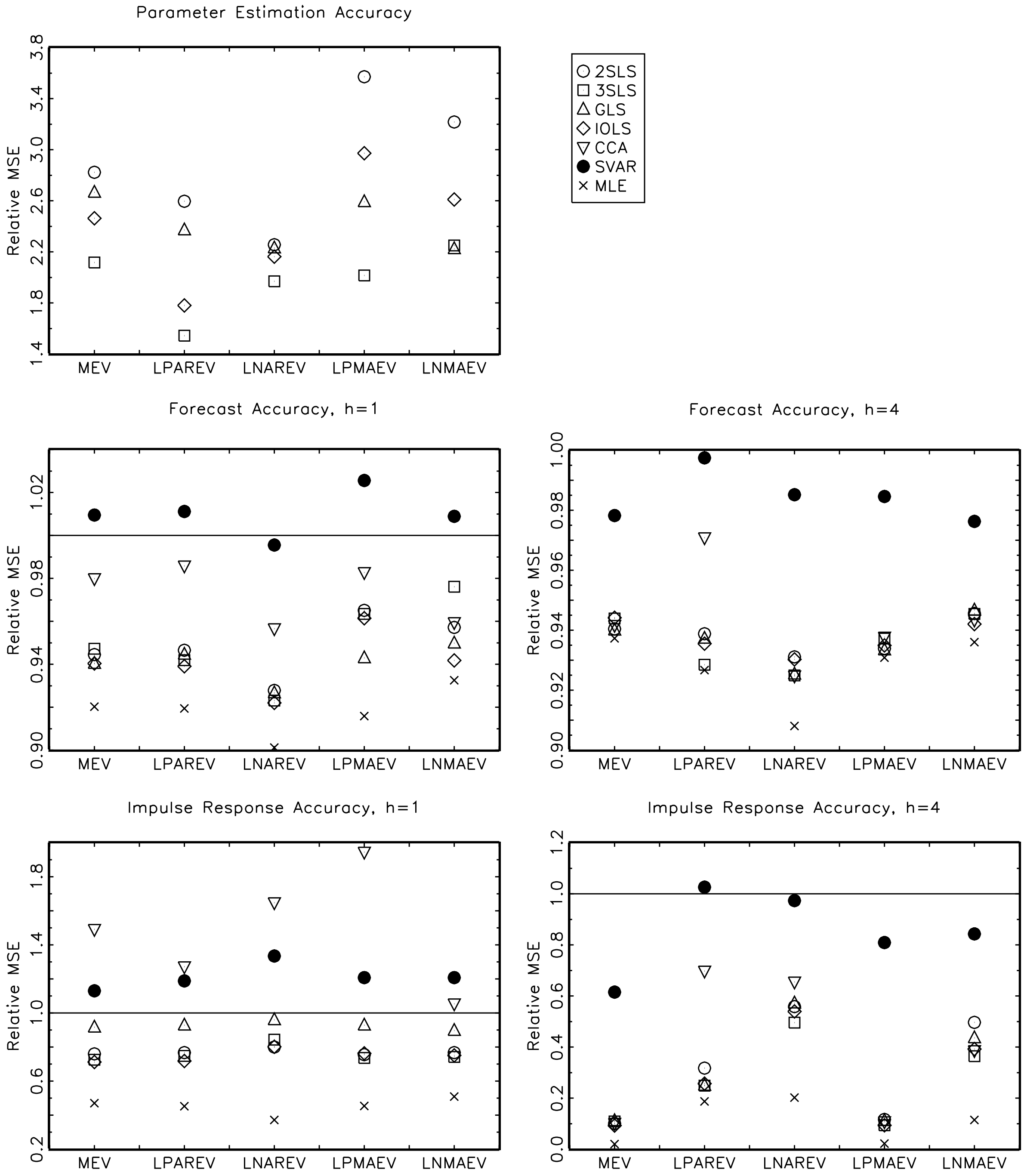

Figure 3.3: MSE ratios for DGP II with $T = 100$.

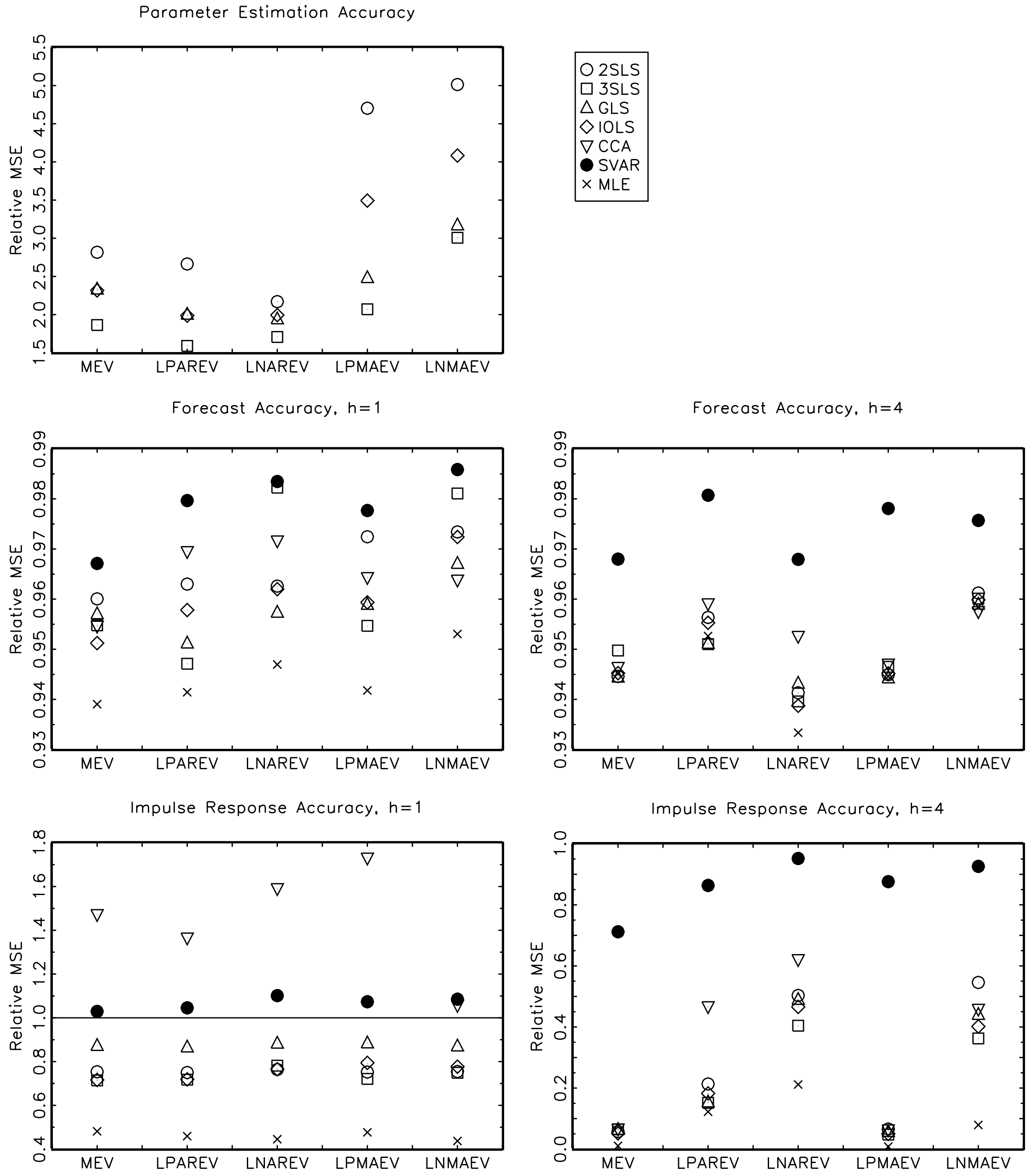

Figure 3.4: MSE ratios for DGP II with $T = 200$.

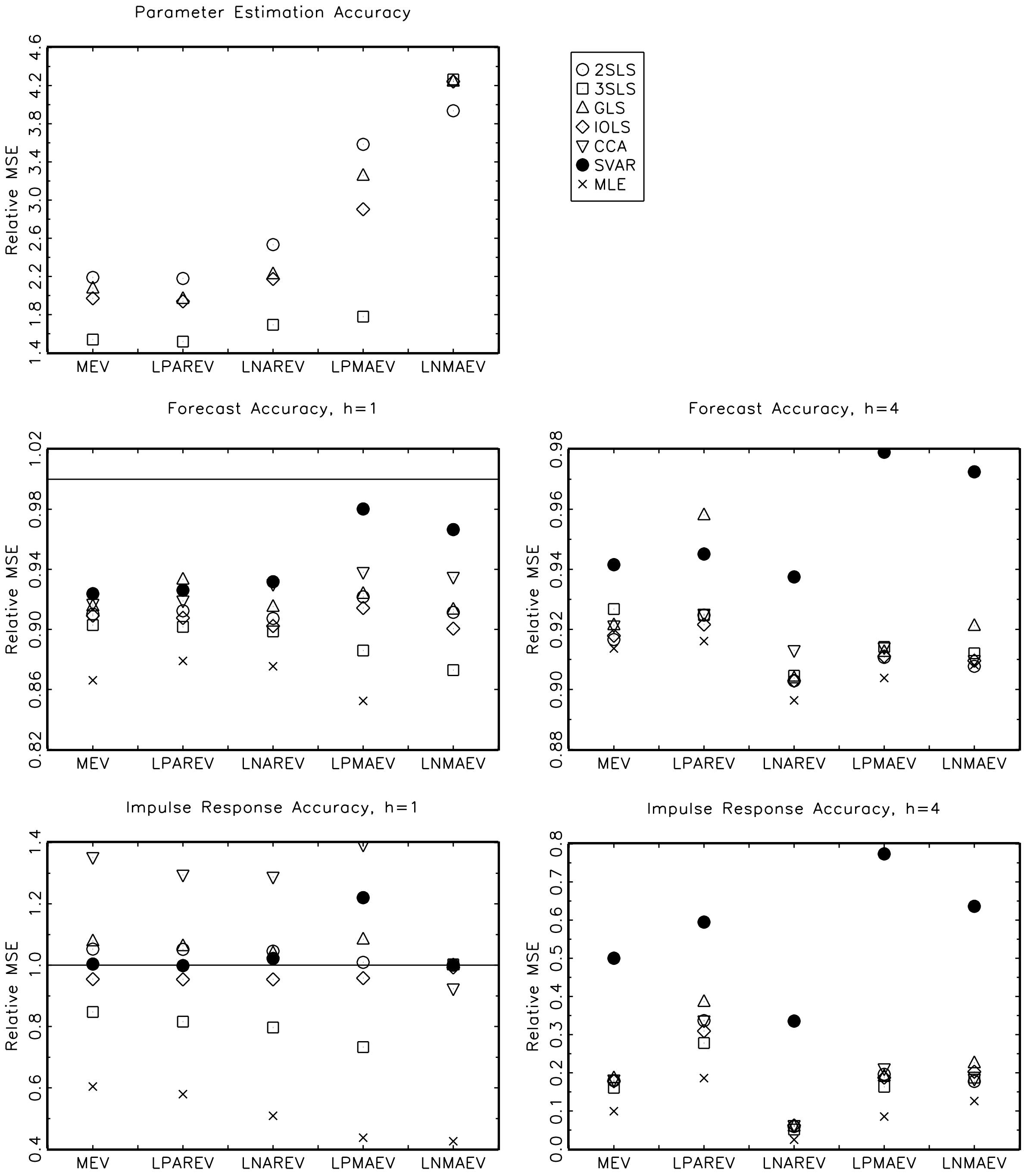

Figure 3.5: MSE ratios for DGP III with $T = 100$.

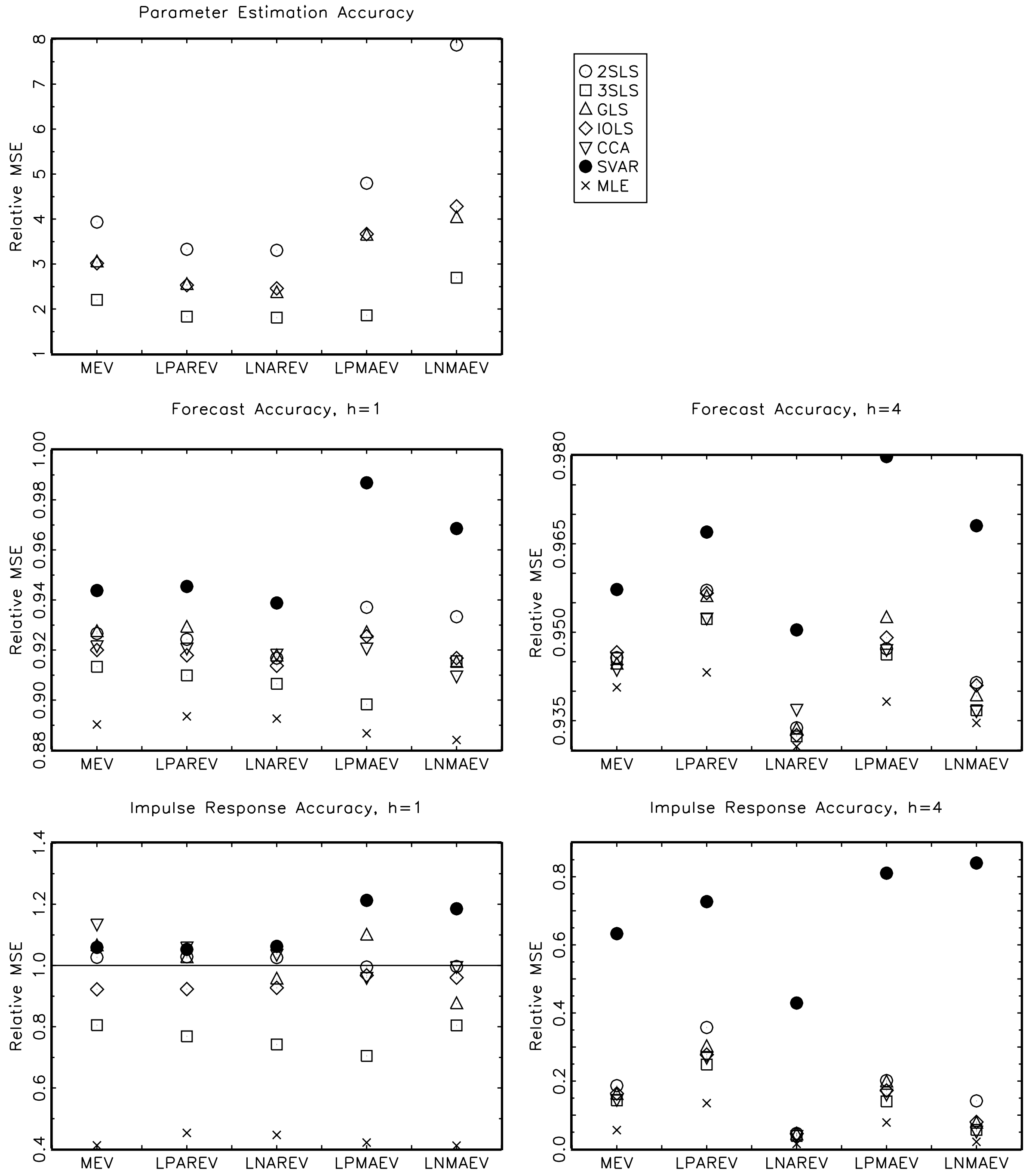

Figure 3.6: MSE ratios for DGP III with $T = 200$.

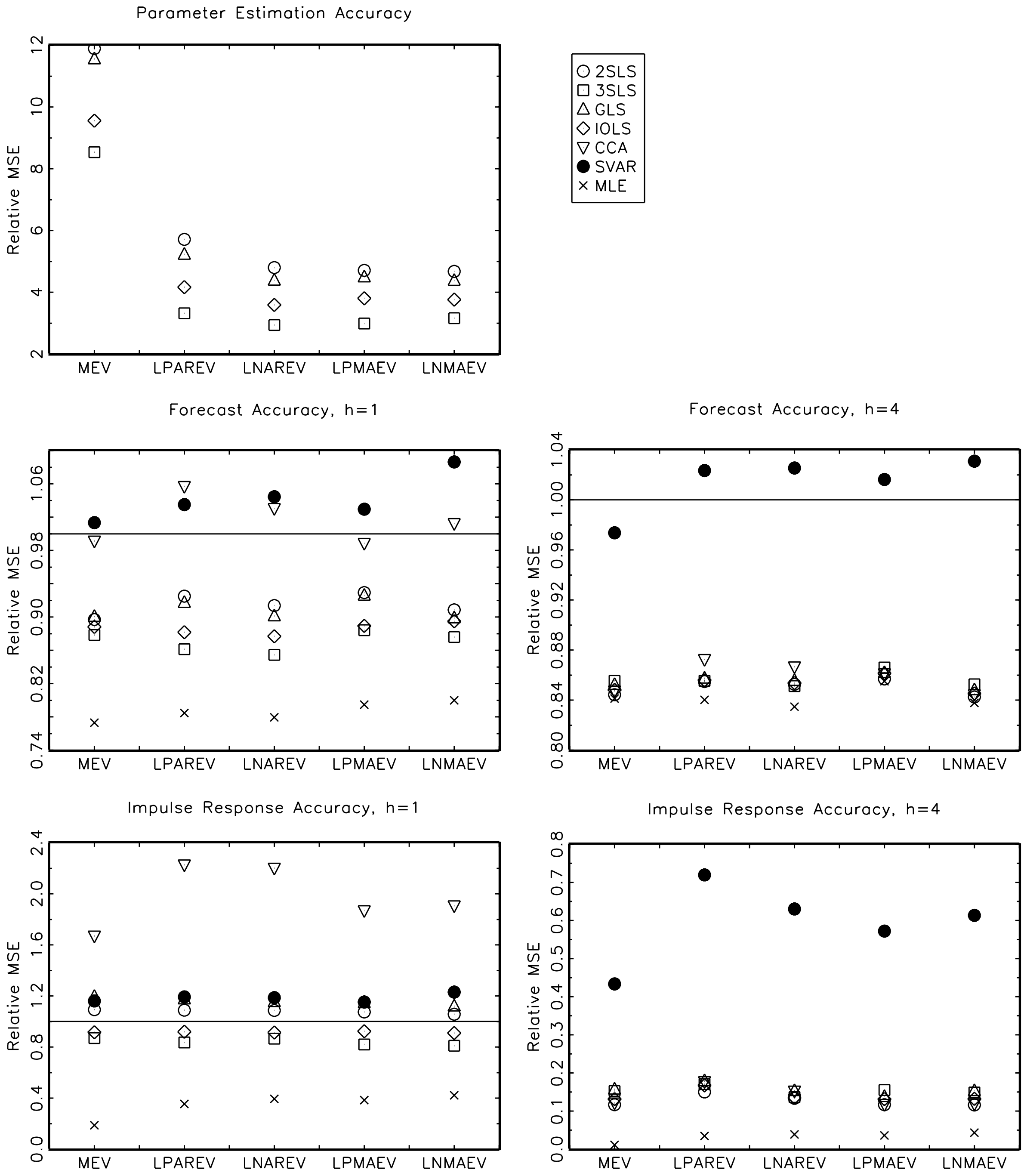

Figure 3.7: MSE ratios for DGP IV with $T = 100$.

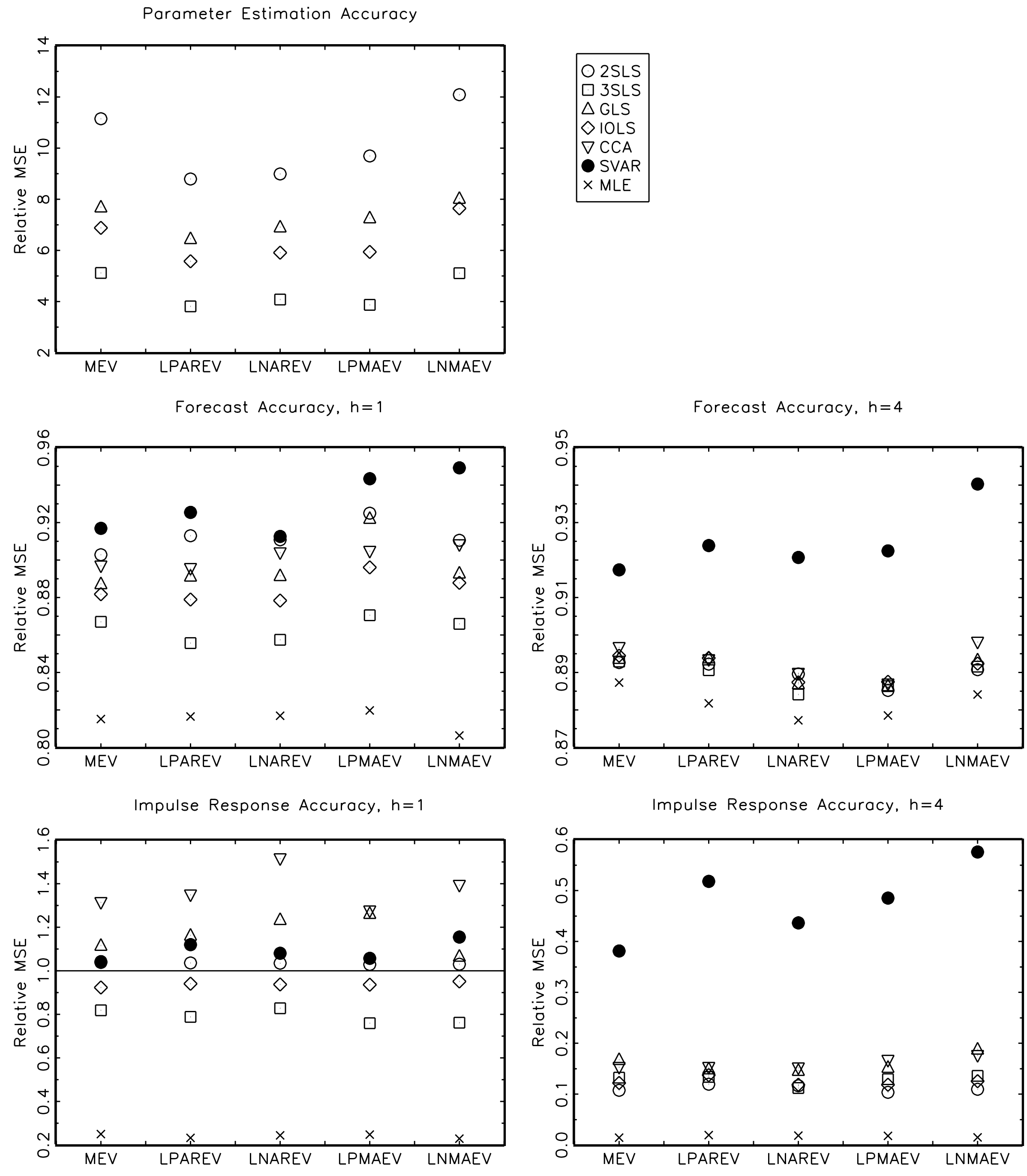

Figure 3.8: MSE ratios for DGP IV with $T = 200$.

Bibliography

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. and Dwyer, M. (1998), 'Business cycle analysis without much theory. A look at structural VARs', *Journal of Econometrics* **83**, 57–88.

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

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

Dufour, J.-M. and 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. and 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. and Serrano, G. R. (2002), 'A Generalized Least Squares Estimation Method For VARMA Models', *Statistics* **13**(4), 303–316.

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

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

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

Hillmer, S. C. and 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. and Salau, M. (2003), 'Generalized Least Squares Estimation of ARMA Models', *Journal of Time Series Anaysis* **24**(2), 165–172.

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

Koreisha, S. and 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. and 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 and P. Dorato, eds, 'Proc. 1983 Amer. Control Conference 2'.

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

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