

**USING SUBSPACE METHODS FOR ESTIMATING ARMA MODELS
FOR MULTIVARIATE TIME SERIES WITH
CONDITIONALLY HETEROSKEDASTIC INNOVATIONS**

**By
Dietmar Bauer**

February 2004

COWLES FOUNDATION DISCUSSION PAPER NO. 1452



**COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY**

**Box 208281
New Haven, Connecticut 06520-8281**

<http://cowles.econ.yale.edu/>

Using subspace methods for estimating ARMA models for multivariate time series with conditionally heteroskedastic innovations

Dietmar Bauer *

Department of Mathematical Methods
in Economics
TU Wien
Argentinierstr. 8, A-1040 Wien

February 17, 2004

Abstract

This paper deals with the estimation of linear dynamic models of the ARMA type for the conditional mean for time series with conditionally heteroskedastic innovation process widely used in modelling financial time series. Estimation is performed using subspace methods which are known to have computational advantages as compared to prediction error methods based on criterion minimization. These advantages are especially strong for high dimensional time series. The subspace methods are shown to provide consistent estimators. Moreover asymptotic equivalence to prediction error estimators in terms of the asymptotic variance is proved. Also order estimation techniques are proposed and analyzed. The estimators are not efficient as they do not model the conditional variance. Nevertheless, they can be used to obtain consistent estimators of the innovations. In a second step these estimated residuals can be used in order to alleviate the problem of specifying the variance model in particular in the multi-output case. This is demonstrated in an ARCH setting, where it is proved that the estimated innovations can be used in place of the true innovations for testing in a linear least squares context in order to specify the structure of the ARCH model without changing the asymptotic distribution.

JEL Classification: C13, C32

Keywords: multivariate models, conditional heteroskedasticity, ARMA systems, subspace methods

*Correspondence to: tel.: ++1 +203 432 6209, e-mail: Dietmar.Bauer@yale.edu

1 Introduction

In financial econometrics it is common practice to use ARMA models for modelling the conditional mean in combination with GARCH type assumptions to allow for conditional heteroskedasticity (see e.g. the references in [Gourieroux, 1997](#)). This is in particular so for high frequency data, where significant autocorrelations in the return series have been observed (see e.g. [Baillie and Bollerslev, 1990](#), and the references contained therein). Most of the studies focus on univariate time series, which is surprising since one of the main applications of models for financial time series lies in aiding the portfolio selection process which unavoidably deals with multivariate time series. Hence techniques for the modeling and estimation of the returns and the variability of several stocks jointly seem to be a relevant issue. While for the return series predominantly ARMA type of models are applied, the models for the conditional variances are legion. This is especially true for multivariate data sets. One of the main reasons for this seems to be the fact that a naive extension of the models for the conditional variance from the univariate case to the multivariate case includes (beside other problems related to the positiveness of the conditional variance) too many parameters for estimation purposes even for large data sets. Assuming the output to be s dimensional the number of parameters for an ARMA model for the vector of returns is linear in s (i.e. for fixed model order, say McMillan degree n). However even a simple (unrestricted) ARCH(1) model for the conditional variance contains more than $s^4/4$ parameters. Beside other problems with multivariate variance models this seems to be the reason for the growing amount of proposed restricted models (see [Gourieroux, 1997](#), Chapter 6 for a somewhat outdated overview). Estimation of the models is usually performed using a three step procedure (see e.g. [Gourieroux, 1997](#), Section 4.2.):

1. Estimate an initial model for the return of the series neglecting the heteroskedasticity. This also leads to estimates of the innovations.
2. Specify and estimate a model for the conditional variances based on a transformation of the estimated innovations (usually either squared innovations or logarithm thereof).
3. Re-estimate the full model including the model for the return and the model for the variance using quasi maximum likelihood based on an assumed distribution (usually Gaussian or student-t) of the standardized innovations.

Such a procedure for the univariate case is discussed in detail in [Mills \(1994, chapter 4\)](#). The third step is often referred to as the BHHH procedure ([Berndt *et al.*, 1974](#)). The asymptotic properties, i.e. consistency and asymptotic normality, of the pseudo ML estimates for the parameter of the model for the conditional mean disregarding the conditional heteroskedasticity are derived in ([Hannan and Deistler, 1988](#), Theorem 4.3.1) under the assumptions of this paper presented below. In the univariate case consistency and asymptotic normality of the pseudo ML estimates under ARCH errors of the full parameter vector including parameters for the variance model has been derived by [Weiss \(1986\)](#). [Jeantheau \(1998\)](#) proved consistency for the full parameter vector for pseudo ML estimators for multivariate ARCH processes¹. Asymptotic normality results for the full parameter vector under restricted assumptions on the model for the return appear to be included in ([Boussama, 1998](#)). For BEKK

¹In fact [Jeantheau \(1998\)](#) contains a more general result dealing with linear filters subject to certain boundedness restrictions both for the model of the conditional mean and the conditional variance.

type of models where the innovations are assumed to be known Comte and Lieberman (2003) prove consistency and asymptotic normality of the pseudo ML estimator of the model for the conditional variance. Ling and McAleer (2003) deal with ARMA-GARCH systems, where the GARCH model is restricted such that conditional correlations are constant over time. They provide consistency and asymptotic normality of the pseudo ML estimators for the full parameter vector including both the model for the conditional mean and the conditional variance.

This paper deals with numerically fast methods for the first step in the above outline. Subspace methods seem to be of interest here, since in the framework considered in this paper they are much faster than traditional prediction error methods based on minimizing the sum of squared innovation estimates (see section 5 for simulations in this respect). This is especially true for high dimensional data sets with, say, tens of outputs. As will be shown in the main part of this paper one does not pay a price in terms of asymptotic accuracy for the numerical advantages.

Subspace methods appeared in the engineering community in the early eighties by the proposal² of CCA in (Larimore, 1983) and became popular in the nineties (Van Overschee and DeMoor, 1994; Verhaegen, 1994). They have been extended to a large number of different model classes (see e.g. Bauer, 2003, for a recent survey). The asymptotic properties for the particular method used in this paper have been analyzed thoroughly in the stationary conditionally homoskedastic case (see again Bauer, 2003, for a survey). The bottom line of these results is that in the case of known order of the true system generating the data the estimates obtained by using CCA are consistent, asymptotically normal and asymptotically equivalent to estimates obtained by minimizing the one step ahead prediction error in the sense that the asymptotic distributions coincide (see section 2 for precise statements).

The aim of this paper is twofold: Firstly as noted above the asymptotic properties for subspace methods have been derived for the conditionally homoskedastic case and showed some analogy to the prediction error methods. Since the asymptotic theory for prediction error estimators has been developed also in the conditionally heteroskedastic case an analogous extension for subspace algorithms seems to be interesting. Secondly and more importantly subspace algorithms do not suffer from numerical problems involved in estimating models using pseudo likelihood maximization or prediction error minimization for time series with many outputs and/or large data sets. The arguably most interesting data sets in financial time series contain a large portfolio of stocks combined with high frequency of measurements (typically five minute returns are used). This paper tries to elucidate the potential of subspace methods for the analysis of such data sets.

The paper is organized as follows: In the next section the model set and the assumptions are discussed. Section 3 presents the CCA subspace method. Section 4 presents the main results of this paper, which are proved in the appendix. Section 5 presents a simulation study. Finally section 6 presents the conclusions of the paper.

Throughout the paper we will use the notation $F_T = o(g_T)$ for random matrix sequences F_T and scalar sequences g_T meaning that $\max_{i,j} |F_{T,i,j}/g_T| \rightarrow 0$ a.s. Here $F_{T,i,j}$ denotes the (i, j) entry of the matrix F_T . Further $F_T = O(g_T)$ means that there exists a constant $C < \infty$ such that $\limsup_{T \rightarrow \infty} \max_{i,j} |F_{T,i,j}/g_T| \leq C$ a.s. Let $o_P(g_T)$ and $O_P(g_T)$ denote the corresponding in probability versions. Note that these concepts differ from the usual definition in situations where the dimension of the matrix F_T depends on T . We assume uniform boundedness of the

²Originally the algorithm has been proposed under the name CVA.

entries rather than boundedness of the norm of the matrix. $\lambda_{max}(A)$ denotes an eigenvalue of maximum modulus of the matrix A . The Kronecker product is denoted by \otimes . The Euclidean norm of a matrix or a vector will be denoted by $\|\cdot\|$.

2 Model set and assumptions

In this paper we consider processes $(y_t)_{t \in \mathbb{Z}}$ generated by a linear, time invariant, finite dimensional, discrete time, state space system of the form

$$\begin{aligned} x_{t+1} &= Ax_t + K\varepsilon_t \\ y_t &= Cx_t + \varepsilon_t + d_t \end{aligned} \tag{1}$$

for $t \in \mathbb{Z}$ where $y_t \in \mathbb{R}^s$ denotes the s -dimensional observed output, $(x_t)_{t \in \mathbb{Z}}$ the unobserved n -dimensional state process and $(\varepsilon_t)_{t \in \mathbb{Z}}$ the s -dimensional innovation sequence. The assumptions on $(d_t)_{t \in \mathbb{Z}}$ will vary in the theorems to follow. Throughout we require $d_t = DT_t$ for some observed strictly stationary process $(T_t)_{t \in \mathbb{Z}}$ having finite fourth moments. Furthermore it is assumed that $(T_t)_{t \in \mathbb{Z}}$ and $(\varepsilon_t)_{t \in \mathbb{Z}}$ are independent. These assumptions include e.g. the constant or cyclical components as well as certain dummy variables (with random timing). Note, however, that e.g. linear time trends (i.e. $T_t = a + bt$) are excluded. $A \in \mathbb{R}^{n \times n}$, $K \in \mathbb{R}^{n \times s}$, $C \in \mathbb{R}^{s \times n}$ are real matrices. Throughout the paper it is assumed that the system is stable, i.e. all the eigenvalues of A are assumed to lie within the open unit disc, and strictly minimum-phase, i.e. all the eigenvalues of $A - KC$ are assumed to lie within the unit circle. Under these assumptions stationary innovation sequences generate stationary output processes $(y_t)_{t \in \mathbb{Z}}$.

It is well known (cf. e.g. Hannan and Deistler, 1988, Chapter 1) that in the given context state space models and ARMA models are just two representations of the same mathematical object, namely the transfer function: It is easy to verify (using the assumptions on the noise sequence $(\varepsilon_t)_{t \in \mathbb{Z}}$ given below) that the stationary solution to the difference equation given above is of the form

$$y_t = \varepsilon_t + \sum_{j=1}^{\infty} K(j)\varepsilon_{t-j} + d_t,$$

where $K(j) = CA^{j-1}K$, $j > 0$ and the infinite sum corresponds to a.s. convergence. The transfer function $k(z)$ describing the input/output mapping, where z as usual denotes the backward shift operator, then is defined as $k(z) = I + zC(I - zA)^{-1}K = I + \sum_{j=1}^{\infty} K(j)z^j$. In the following we will provide a very brief discussion of the main concepts in the state space framework that are needed for the estimation theory in this paper. The presentation of the concepts is intended to serve as a reference list of technical terms in the state space framework and is hence far from being self-contained. For a detailed discussion we refer the interested reader to Chapters 1 and 2 of Hannan and Deistler (1988).

$k(z) = I + zC(I - zA)^{-1}K$ is a matrix valued function which is rational in z seen as a complex variable. Therefore the transfer function has a representation as an ARMA system according to a left matrix fraction representation $k(z) = a^{-1}(z)b(z)$, where $a(z) = I + a_1z + \dots + a_pz^p$, $b(z) = I + b_1z + \dots + b_qz^q$. Furthermore $(y_t - d_t)_{t \in \mathbb{Z}}$ satisfies the corresponding ARMA vector difference equations $a(z)y_t = b(z)\varepsilon_t$. Conversely each ARMA process has a representation as a solution to state space equations. See (Hannan and Deistler, 1988, p. 15) for an explicit construction of a state space system based on the ARMA representation.

In this sense ARMA representations and state space representations are equivalent. A more detailed discussion on the relation between ARMA and state space systems can be found in (Hannan and Deistler, 1988, section 1.2).

Let π denote the mapping attaching the transfer function $k(z) = I + zC(I - zA)^{-1}K$ to the state space system (A, K, C) . State space systems corresponding to a given transfer function $k(z)$ are not unique, i.e. the mapping π is not injective. There are two sources of nonuniqueness: The choice of the state basis and non-minimality. A state space system (A, K, C) is called *minimal*, if no other state space system $(\tilde{A}, \tilde{K}, \tilde{C})$ exists, such that $\pi(A, K, C) = \pi(\tilde{A}, \tilde{K}, \tilde{C})$, where $\tilde{A} \in \mathbb{R}^{\tilde{n} \times \tilde{n}}, A \in \mathbb{R}^{n \times n}$ such that $\tilde{n} < n$. In other words minimality refers to minimal state dimension. When modeling the transfer function $k(z)$ it is no restriction of generality to restrict attention to minimal systems which will be done henceforth. The set of transfer functions corresponding to minimal, stable, minimum-phase state space systems with state dimension equal to n will be denoted as $M(n)$. For $k(z) \in M(n)$ the integer n will be called the *order* of the system.

Two systems (A, K, C) and $(\tilde{A}, \tilde{K}, \tilde{C})$ are called *observationally equivalent* if $\pi(A, K, C) = \pi(\tilde{A}, \tilde{K}, \tilde{C})$. The set of minimal observationally equivalent systems can be described using the group of nonsingular matrices $T \in \mathbb{R}^{n \times n}$: Two minimal systems (A, K, C) and $(\tilde{A}, \tilde{K}, \tilde{C})$ are observationally equivalent if and only if there exists a nonsingular transformation T of the state basis such that $A = T\tilde{A}T^{-1}, K = T\tilde{K}, C = \tilde{C}T^{-1}$ corresponding to $x_t = T\tilde{x}_t$. Therefore minimality is not restrictive enough in order to uniquely specify one state space system corresponding to a particular transfer function $k(z)$. Such a one-to-one relation is usually used in order to define a parameterization of subsets of $M(n)$. In this paper we will use the overlapping echelon forms presented in section 2.6 of Hannan and Deistler (1988) to define a parameterization of $M(n)$. Overlapping forms make it possible to parameterize continuously generic (and hence not disjoint) pieces of $M(n)$. In general overlapping forms are defined as a collection of bijective mappings $\varphi_i : T_i \subset \mathbb{R}^{2ns} \rightarrow M(n; i), i \in I$, where I is a finite index set, such that each set $M(n; i)$ is open and dense in $M(n)$ and each $k(z) \in M(n)$ lies in the interior of some $M(n; i)$. The mappings φ_i are continuous where the topology in $M(n)$ is the so called pointwise topology and T_i is equipped with the Euclidean topology. For details on the structure of the various pieces $M(n; i)$ of $M(n)$ and the geometrical properties of the decomposition see section 2.6 in Hannan and Deistler (1988). For this paper it is sufficient to state that to each $\theta \in T_i$ the parameterization attaches a system $(A(\theta), K(\theta), C(\theta))$ such that $k(z, \theta) = \pi(A(\theta), K(\theta), C(\theta)) \in M(n; i)$. Conversely for each $k(z) \in M(n)$ there exists for each index i such that $k(z) \in M(n; i)$ a unique parameter vector $\theta_i = \varphi_i^{-1}(k(z))$ in the interior of T_i . Furthermore the mapping $\rho_i : \theta \mapsto (A(\theta), K(\theta), C(\theta))$ is differentiable for $\theta \in T_i$. Finally T_i is an open subset of \mathbb{R}^{2ns} .

Throughout this paper we will always use the following assumptions on the noise:

Assumption 1 *The process $(\varepsilon_t)_{t \in \mathbb{Z}}$ is assumed to be an ergodic, strictly stationary, martingale difference sequence with respect to the sequence of increasing sigma fields $\mathcal{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \dots\}$ having the following properties:*

$$\begin{aligned} \mathbb{E}\{\varepsilon_t | \mathcal{F}_{t-1}\} &= 0 \quad , \quad \lim_{k \rightarrow \infty} \mathbb{E}\{\varepsilon_t \varepsilon'_t | \mathcal{F}_{t-k}\} = \Omega = \mathbb{E}\varepsilon_t \varepsilon'_t, a.s. \\ \mathbb{E}\varepsilon_{t,j}^4 &< \infty \quad , \quad j = 1, \dots, s, \end{aligned}$$

where $\varepsilon_{t,i}$ denotes the i -th component of the vector ε_t .

These assumptions on the innovation sequence hold for a number of commonly used models. Among these are:

- univariate ARCH(p) processes (Engle, 1982): These processes are defined by the equations

$$\varepsilon_t = \eta_t \sqrt{h_t}, \quad h_t = c + \sum_{j=1}^p a_j \varepsilon_{t-j}^2,$$

where $(\eta_t)_{t \in \mathbb{Z}}$ is i.i.d. standard normally distributed. It follows that $h_t = \mathbb{E}\{\varepsilon_t^2 \mid \mathcal{F}_{t-1}\}$. The solutions to these difference equations are stationary, if $a_j \geq 0, c > 0, \sum_{j=1}^p a_j < 1$. Bougerol and Picard (1992) show that in this situation the solutions process is also ergodic. Defining the matrix $\psi \in \mathbb{R}^{p \times p}$ whose (i, j) entry is equal to $\psi_{i,j} = a_{i+j} + a_{i-j}, i, j = 1, \dots, p$, where $a_j = 0, j < 1$ or $j > p$ is used, it follows that the fourth moment is finite, if $3(a_1, \dots, a_p)(I - \psi)^{-1}(a_1, \dots, a_p)' < 1$ (cf. Gouriéroux, 1997, Exercise 3.4). The derivation of this result depends on Gaussianity of η_t . The condition $\lim_{k \rightarrow \infty} \mathbb{E}\{\varepsilon_t^2 \mid \mathcal{F}_{t-k}\} = \mathbb{E}\varepsilon_t^2$ follows from the proof of ergodicity and strict stationarity given in Bougerol and Picard (1992).

- univariate GARCH(p,q) processes (Bollerslev, 1986): Here

$$h_t = c + \sum_{j=1}^p a_j \varepsilon_{t-j}^2 + \sum_{j=1}^q b_j h_{t-j}.$$

Again for $\varepsilon_t = \eta_t \sqrt{h_t}$, where $(\eta_t)_{t \in \mathbb{Z}}$ is i.i.d. standard normally distributed Bougerol and Picard (1992) show that the generated process is strictly stationary and ergodic for $c > 0, a_j \geq 0, b_j \geq 0, \sum_{j=1}^p a_j + \sum_{j=1}^q b_j < 1$. Conditions for the existence of a fourth moment are derived in Ling and McAleer (2002) without the assumption of normality of η_t . The condition on the conditional second moments again follows from Bougerol and Picard (1992).

- E-GARCH processes (Nelson, 1991): Here

$$\log h_t = \alpha + \sum_{j=1}^{\infty} \beta_j g(\eta_{t-j}), \quad g(z) = \theta z + \gamma(|z| - \mathbb{E}|z|),$$

and the normalized innovations $(\eta_t)_{t \in \mathbb{Z}}$ are assumed to be i.i.d. distributed according to a GED type of distribution with tail thickness parameter $\nu > 1$ with mean zero and unit variance. Also for the solution processes to these difference equations the assumptions 1 are fulfilled (cf. Nelson, 1991).

- BEKK-ARCH (Engle and Kroner, 1995): The process is defined as $\varepsilon_t = H_t^{1/2} \eta_t$, where $(\eta_t)_{t \in \mathbb{Z}}$ is i.i.d. distributed with zero mean, unit variance and everywhere positive continuous density.

$$H_t = H_0 + \sum_{i=1}^m \sum_{j=1}^n A_{ij} \varepsilon_{t-i} \varepsilon_{t-i}' A_{ij}', \quad H_0 = H_0' > 0.$$

Here $H_t^{1/2}$ refers to the Cholesky factor of H_t . The number of free parameters contained in this model is equal to $mns^2 + s(s+1)/2$. Rahbek *et al.* (2003) present an analysis of the stochastic properties of the solution processes based on Markov chain theory. Let

$$\Phi = \begin{bmatrix} \sum_{j=1}^n A_{1j} \otimes A_{1j}, \dots, \sum_{j=1}^n A_{m-1j} \otimes A_{m-1j}, & \sum_{j=1}^n A_{mj} \otimes A_{mj}, \\ I_{s(m-1)} \otimes I_{s(m-1)}, & 0_{s^2(m-1) \times s^2} \end{bmatrix} \in \mathbb{R}^{ms^2 \times ms^2}.$$

If $|\lambda_{max}(\Phi)| < 1$ then there exists a solution process $(\varepsilon_t)_{t \in \mathbb{Z}}$ that is strictly stationary, ergodic and has finite second moments. Furthermore the solution is even geometrically ergodic and hence the conditional second moment tends to the unconditional for conditioning horizon tending to infinity. Therefore assumption 1 holds for this class of processes if the fourth moment is finite. If further $m = n = 1$ then $|\lambda_{max}(\Phi)| < 1/\sqrt{3}$ is a sufficient condition for the fourth moments to be finite. In the univariate case the same condition as in the ARCH(1) model cited above is obtained.

Rahbek *et al.* (2003) also discuss a number of different model classes, which will not be cited here. Interestingly enough in the list above multivariate ARCH is not included. In the multivariate case the requirement of H_t to be a covariance matrix seems to complicate the formulation of the model considerably. Note that in this paper we do not assume any particular model structure for the conditional variance, but use the high level assumptions 1. Another remark is in order: As a main application we referred to high frequency data sets. It has been observed that in these data sets the volatility does not only show conditional heteroskedasticity in the form of dependence on the return history but also a cyclical component related to the time of the day. This additional heteroskedasticity component is sometimes modelled as a time dependent multiplicative constant contradicting the strict stationarity assumption. However, if one collects in this situation all returns within a given day into a large vector the corresponding time series remains in the scope of the present paper. Obviously the time series using the high frequent time scale as increments and the time series using one day as the time increment are in one-to-one correspondence and results derived in the one framework typically can be translated to the other framework. In particular the sample covariance sequence of the high frequency data is just a finite summation of the entries of the daily sample covariance sequence. Therefore the results on estimation accuracy can easily be transferred from the stationary daily framework to the nonstationary high frequency framework.

The main problem dealt with in this paper is the estimation of the state space model for the conditional mean based on observations y_1, \dots, y_T of the output.

3 Estimation methods

The traditional method of estimation is pseudo maximum likelihood ³ which is described below. Prior to estimation of the model for the mean, deterministic terms will be dealt with typically. These terms have been included as d_t where $d_t = DT_t$ for some matrix D and some observed process T_t has been assumed. Let \hat{y}_t denote the residuals of a regression of y_t onto T_t and let

$$\varepsilon_t(\theta) = \hat{y}_t - \sum_{j=1}^{t-1} C(\theta)(A(\theta) - K(\theta)C(\theta))^{j-1}K(\theta)\hat{y}_{t-j}$$

denote the one step ahead prediction error based on the assumption ⁴ $x_1 = 0$ at the parameter vector $\theta \in T_i$. Assume that a constant conditional variance $\Omega = \Omega(\omega)$ is parameterized using the parameter vector $\omega \in S \subset \mathbb{R}^{s(s+1)/2}$ and the true innovation $\varepsilon_t = \Omega^{1/2}(\omega_0)\eta_t$, where

³Also the name quasi maximum likelihood is sometimes used in the literature.

⁴(Hannan and Deistler, 1988) show that under the stability and strict minimum-phase assumption the choice of the initial values is not essential for the asymptotic properties of the estimators.

η_t is conditionally standard normally distributed. Then $-2/T$ times the logarithm of the likelihood is equal to

$$L(\hat{y}_t, \theta, \omega) = \log \det \Omega(\omega) + \left(\frac{1}{T} \sum_{t=1}^T \varepsilon_t(\theta)' \Omega(\omega)^{-1} \varepsilon_t(\theta) \right).$$

The pseudo ML estimator then is obtained as the minimizing argument of this function, i.e.

$$\tilde{\tau}_i = \arg \min_{\tau \in \Psi_i} L(\hat{y}_t, \tau)$$

where $\tau = [\theta', \omega']' \in \Psi_i \subset T_i \times S$. Note that here we use a specific piece $M(n; i)$ of $M(n)$ corresponding to the index i for the optimization rather than optimizing over the full $M(n)$. The fact that $M(n; i)$ is open and dense in $M(n)$ justifies this choice assuming that i is chosen such that the true transfer function $k(z)$ is contained in the interior of $M(n; i)$ which will implicitly always be assumed. Then it follows from the consistency results in section 4.2. of Hannan and Deistler (1988) that it is no restriction of generality to assume that the pseudo ML estimator also lies in $M(n; i)$ for T large enough. Numerically the estimator is obtained by using gradient search methods starting at an initial guess. Here the differentiability property of the parameterization noted in section 2 is essential. In the (likely) case that $(\eta_t)_{t \in \mathbb{Z}}$ is not i.i.d. standard normally distributed or the model for the conditional variance is misspecified the function $L(\hat{y}_t, \theta, \omega)$ might still be a reasonable criterion function leading to estimators having desirable asymptotic properties as documented by the results in Chapter 4 of Hannan and Deistler (1988).

For $\tau' = [\theta', \omega']$ where θ parameterizes the model for the return and ω corresponds to a parameterization for Ω and there are no restrictions on ω and no cross restrictions, i.e. $\Psi_i = T_i \times S$, then the criterion function can be concentrated with respect to ω leading to the criterion function

$$L(\hat{y}_t, \theta, \hat{\omega}) = \log \det \left(\frac{1}{T} \sum_{t=1}^T \varepsilon_t(\theta) \varepsilon_t(\theta)' \right).$$

The minimizing parameter vector $\hat{\theta}_i$ is called the prediction error estimate ⁵. If the true data generating process is in fact conditionally heteroskedastic, but the innovations fulfill assumptions 1, then the theory of Hannan and Deistler (1988), Chapter 4, still applies and the corresponding estimator $\hat{\theta}_i$ is consistent and asymptotically normal. For Gaussian homoskedastic innovations the prediction error estimator attains the Cramer Rao lower bound and hence is asymptotically efficient. For conditionally heteroskedastic innovations, however, it loses the efficiency property.

Note that the treatment of the deterministic term DT_t is nonstandard since it is not included in the criterion function but rather assumed to be estimated prior to pseudo maximum likelihood estimation of the parameter τ . This will not lead to identical estimates. It follows from standard evaluations that in the case where each component $T_{t,i}$ of T_t is a harmonic process of the form

$$T_{t,i} = \mathcal{R}(\mu_i^t T_{0,i}^c)$$

for $t \in \mathbb{Z}$ for some $\mu_i \in \mathbb{C}, |\mu_i| = 1, T_{0,i}^c \in \mathbb{C}$ where \mathcal{R} denotes the real part of the variable, the asymptotic distribution of the estimates \hat{D} is not altered whether it is estimated using

⁵In the literature also estimators minimizing $\text{tr}[\Sigma \frac{1}{T} \sum_{t=1}^T \varepsilon_t(\theta) \varepsilon_t(\theta)']$ are termed prediction error estimator. We will not use this terminology.

least squares fitting prior to pseudo likelihood minimization or the term DT_t is included in the likelihood. Furthermore if in that case $\hat{\theta}$ denotes the prediction error estimator based on \hat{y}_t and $\tilde{\theta}$ the (unfeasible) estimator based on $y_t - DT_t$ then $\sqrt{T}(\hat{\theta} - \tilde{\theta}) = o_P(1)$.

Pseudo maximum likelihood estimation as presented above results in a nonlinear optimization problem which in particular for high dimensional parameter sets is numerically problematic (see e.g. the simulations in section 5). The search for numerically well behaved procedures has been the motivation for the introduction of CCA in Larimore (1983). The CCA algorithm uses the properties of the state in order to obtain estimates of the system matrices. In this section we will not distinguish between y_t, \tilde{y}_t and \hat{y}_t notationally. For reasons of readability we will only use y_t under the implicit assumption that $D = 0$, whereas the estimation of course can be performed on the basis of \hat{y}_t . The state is not observed directly. However, on the basis of the system equations (1) the state can be reconstructed from the past of the time series using the recursions defining the state:

$$\begin{aligned} x_t &= Ax_{t-1} + K\varepsilon_{t-1} = Ax_{t-1} + K(y_{t-1} - Cx_{t-1}) \\ &= (A - KC)x_{t-1} + Ky_{t-1} = (A - KC)^2x_{t-2} + Ky_{t-1} + (A - KC)Ky_{t-2} \\ &= \dots = (A - KC)^px_{t-p} + \sum_{j=0}^{p-1} (A - KC)^j Ky_{t-1-j} = \sum_{j=1}^{\infty} (A - KC)^{j-1} Ky_{t-j} \end{aligned}$$

using the strict minimum-phase assumption implying $(A - KC)^p \rightarrow 0$ for $p \rightarrow \infty$. Again convergence in the infinite sum is a.s. Defining $\mathcal{K}_p = [K, (A - KC)K, \dots, (A - KC)^{p-1}K]$, $Y_{t,p}^- = [y'_{t-1}, y'_{t-2}, \dots, y'_{t-p}]'$ we obtain

$$x_t = (A - KC)^px_{t-p} + \mathcal{K}_p Y_{t,p}^-.$$

Hence the state lies in the space spanned by the past of y_t , i.e. $\text{span}\{y_{k,i}, k < t, i = 1, \dots, s\}$. Consider the best mean square prediction, $y_{t+j|t-1}$ say, of y_{t+j} , $j \geq 0$ based on $y_s, s < t$. It follows from the system equations that

$$y_{t+j} = Cx_{t+j} + \varepsilon_{t+j} = CA^j x_t + \varepsilon_{t+j} + \sum_{i=1}^j CA^{i-1} K \varepsilon_{t+j-i}.$$

Since $(\varepsilon_t)_{t \in \mathbb{Z}}$ is assumed to be a martingale difference we obtain $y_{t+j|t-1} = CA^j x_t, j \geq 0$. Therefore the prediction for all horizons is a linear function of the state. In this sense the state contains all information of the past relevant for predicting the future.

These two facts can be combined into the following central equation by writing the equation jointly for $j = 0, \dots, f - 1$ and $t \in \mathbb{Z}$

$$Y_{t,f}^+ = \mathcal{O}_f x_t + \mathcal{E}_f E_{t,f}^+ = \mathcal{O}_f \mathcal{K}_p Y_{t,p}^- + \mathcal{O}_f (A - KC)^p x_{t-p} + \mathcal{E}_f E_{t,f}^+,$$

where $Y_{t,f}^+ = [y'_t, y'_{t+1}, \dots, y'_{t+f-1}]'$ and $E_{t,f}^+$ is defined analogously by using ε_t instead of y_t . Further $\mathcal{O}_f = [C', A'C', \dots, (A')^{f-1}C']'$ and \mathcal{E}_f denotes the matrix whose i -th block row for $i > 1$ is equal to $[CA^{i-2}K, \dots, CK, I, 0, \dots, 0]$, while the first block row is equal to $[I, 0, \dots, 0]$. Note that this equation decomposes $Y_{t,f}^+$ into three parts:

1. $\mathcal{O}_f \mathcal{K}_p Y_{t,p}^-$, where typically f and p are selected large enough such that $\mathcal{O}_f \mathcal{K}_p$ is rank reduced.

2. $\mathcal{O}_f(A-KC)^p x_{t-p}$ will be small for large p due to the strict minimum-phase assumption.
3. $\mathcal{E}_f E_{t,f}^+$ is uncorrelated with the remaining terms.

This motivates a class of estimation algorithms which can be described as follows:

1. Obtain a rank n matrix $\hat{\beta} = \hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ estimating $\mathcal{O}_f \mathcal{K}_p$ by using rank restricted regression techniques in the equation $Y_{t,f}^+ = \beta Y_{t,p}^- + N_{t,f}^+, t = 1, \dots, T$ where not available observations are replaced by zero. Alternatively the regression can be performed for $t = f+1, \dots, T-p$ without changing the asymptotic results.
2. Estimate the state as $\hat{x}_t = \hat{\mathcal{K}}_p Y_{t,p}^-, t = p+1, \dots, T+1, \hat{x}_t = 0, t = 1, \dots, p$. The system matrix estimates are obtained using least squares fitting in the system equations where the estimate \hat{x}_t of the state is used in place of the true state x_t . Use the notation $\langle a_t, b_t \rangle = T^{-1} \sum_{t=1}^T a_t b_t'$, where we assume that not available data is replaced by zeros. Here we use the same symbol for the processes $(a_t)_{t \in \mathbb{Z}}, (b_t)_{t \in \mathbb{Z}}$ and the random variables a_t and b_t respectively which should not cause any confusion. Then

$$\begin{aligned} \hat{C} &= \langle y_t, \hat{x}_t \rangle \langle \hat{x}_t, \hat{x}_t \rangle^{-1} \quad , \quad \hat{A} = \langle \hat{x}_{t+1}, \hat{x}_t \rangle \langle \hat{x}_t, \hat{x}_t \rangle^{-1}, \\ \hat{\varepsilon}_t &= y_t - \hat{C} \hat{x}_t, t = 1, \dots, T \quad , \quad \hat{K} = \langle \hat{x}_{t+1}, \hat{\varepsilon}_t \rangle \langle \hat{\varepsilon}_t, \hat{\varepsilon}_t \rangle^{-1}. \end{aligned}$$

The class of algorithms described above has been called *subspace algorithms* in the engineering literature. Note, however, that the term subspace algorithm is also used for a different class of algorithms where the center of attention is the estimation of the range space of the observability matrix \mathcal{O}_f , which motivates the term. These procedures have in common that the first step usually is presented as two steps including a least squares fit followed by a model reduction step based on a singular value decomposition (SVD). For a review of both classes of subspace methods see Bauer (2003).

The rank restricted regression in the first step amounts to the minimization of the criterion function

$$\hat{\beta} = \arg \min_{\beta: \text{rank}(\beta)=n} \text{tr} \left[\hat{W} \langle Y_{t,f}^+ - \beta Y_{t,p}^-, Y_{t,f}^+ - \beta Y_{t,p}^- \rangle \right]$$

where $\hat{W} = \hat{W}' > 0$ is a design variable to be chosen by the user. $\hat{W} = \langle Y_{t,f}^+, Y_{t,f}^+ \rangle^{-1}$ corresponds to the maximum likelihood estimate assuming Gaussian i.i.d. errors $N_{t,f}^+$. The algorithm obtained using this choice has been called **CVA** by Larimore (1983) and **CCA** by Deistler *et al.* (1995) where the acronym stands for canonical variate (resp. correlation) analysis. The estimate can be obtained using the singular value decomposition

$$\hat{W}^{1/2} \langle Y_{t,f}^+, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1/2} = \hat{U} \hat{\Sigma} \hat{V}' = \hat{U}_n \hat{\Sigma}_n \hat{V}_n' + \hat{R}_n$$

where $\hat{U}_n \in \mathbb{R}^{sf \times n}$ denotes the matrix of left singular vectors corresponding to the n dominating singular values. $\hat{\Sigma}_n = \text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_n)$ denotes the diagonal matrix containing the dominant n singular values ordered decreasing in size. $\hat{V}_n \in \mathbb{R}^{sp \times n}$ conformably contains the right singular vectors. Here $X^{1/2}$ denotes the uniquely defined symmetric square root of a matrix X . In this step the order of the estimated system has to be supplied.

One possibility to estimate the order is to use the information contained in the estimated singular values. Define the criterion

$$IC(n) = \|\hat{R}_n\|^2 + \frac{d(n)C_T}{T}$$

which can be minimized over $n = 0, 1, \dots, H_T$ for some upper bound $H_T \leq \min(fs, ps)$ to obtain the order estimate. Here $d(n) = 2ns$ denotes the number of parameters in a model of order n and $C_T > 0, C_T \rightarrow \infty, C_T/T \rightarrow 0$ denotes a penalty term. Although the form is similar to the definition of information criteria no relation has been established. Typical choices for $\|\cdot\|$ are the Frobenius norm (Peternell, 1995), the two norm (Bauer, 2001) and restricted to the case of CCA weights $\|\hat{R}_n\|^2 = -\sum_{j=n+1}^M \log(1 - \hat{\sigma}_j^2)$, which is not a norm (Camba-Mendez and Kapetanios, 2001).

An open point in the description of the algorithm is the choice of the integers f and p . In the conditionally homoskedastic stationary case it has been suggested to use $f = p = 2\hat{p}_{AIC}$, where \hat{p}_{AIC} denotes the AIC order estimate in an autoregressive approximation of y_t . Also in the case of heteroskedastic innovations this choice is sufficient for the results of the next section to hold. Throughout the text we will use two different scenarios for the choice of the user parameters:

Assumption 2 f and \hat{W}_f^+ are chosen corresponding to one of the following two possibilities:

- $f \geq n$ fixed and independent of the sample size where n denotes the true order. \hat{W}_f^+ is such that there exists a positive definite matrix W_f^+ where $\hat{W}_f^+ - W_f^+ = O(Q_T)$.
- $f = p$ depending on the sample size and $\hat{W}_f^+ = \langle Y_{t,f}^+, Y_{t,f}^+ \rangle^{-1/2}$. This choice will be labeled CCA in the following.

In any case $p \rightarrow \infty, p = O((\log T)^a)$ for some $a < \infty$.

4 Asymptotic properties

In this section we will state the results of this paper. Let $\tilde{y}_t = y_t - DT_t$ and $\hat{y}_t = y_t - \hat{D}T_t$ where $\hat{D} = \langle y_t, T_t \rangle \langle T_t, T_t \rangle^{-1}$. The key to the results is the uniform convergence result for covariance sequences as stated in (Hannan and Deistler, 1988, Theorem 5.3.2, see also (5.3.7)): Let $\hat{\gamma}_j = \langle \tilde{y}_t, \tilde{y}_{t-j} \rangle$ and $\dot{\gamma}_j = \sum_{i=0}^{\infty} K_{i+j} \dot{\Omega} K_i'$, where $\dot{\Omega} = \langle \varepsilon_t, \varepsilon_t \rangle$. Then (5.3.7) of Hannan and Deistler (1988) states that for $H_T = o((\log T)^a)$ for some $a < \infty$ and $Q_T = \sqrt{\log \log T/T}$ it holds that

$$\max_{|j| \leq H_T} \|\hat{\gamma}_j - \dot{\gamma}_j\| = O(Q_T) \quad \text{a.s.} \quad (2)$$

However, \tilde{y}_t is not observed and has to be replaced by $\hat{y}_t = y_t - \hat{D}T_t$. Corresponding to the process $(T_t)_{t \in \mathbb{Z}}$ we will use the following assumptions:

Assumption 3 The process $(T_t)_{t \in \mathbb{Z}}$ is strictly stationary with finite fourth moments independent of $(\varepsilon_t)_{t \in \mathbb{Z}}$ where either of the following two conditions hold:

- $(T_t)_{t \in \mathbb{Z}}$ is ergodic.
- $T_{t,i} = \mathcal{R}(\mu_i^t T_{0,i}^c)$ where $T_{t,i}$ denotes the i -th component of T_t , $|\mu_i| = 1, \mu_i \in \mathbb{C}, T_{0,i}^c \in \mathbb{C}$ is a random variable.

Under these assumptions it follows that

$$\max_{|j| \leq H_T} \|\langle \tilde{y}_t, \tilde{y}_{t-j} \rangle - \langle \hat{y}_t, \hat{y}_{t-j} \rangle\| = O(Q_T^2) \quad \text{a.s.} \quad (3)$$

and hence the difference is negligible. This is a consequence of $\hat{y}_t = \tilde{y}_t + (D - \hat{D})T_t$ where under the current assumptions $\hat{D} - D = O(Q_T)$. The result then follows from

$$\max_{|j| \leq H_T} \|\langle T_t, \tilde{y}_{t-j} \rangle\| = O(Q_T) \quad \text{a.s.} \quad (4)$$

corresponding to Theorem 5.3.4. of Hannan and Deistler (1988) under ergodicity of $(T_t)_{t \in \mathbb{Z}}$. The result for the second set of assumptions for $(T_t)_{t \in \mathbb{Z}}$ following from the arguments on p. 159 of Hannan and Deistler (1988) and the structure of T_t . Hence the estimate of the covariance sequence converges uniformly in the lag (up to some upper bound) at rate Q_T . Note that $\hat{\gamma}_j - \mathbb{E}\tilde{y}_t\tilde{y}'_{t-j} = \sum_{i=0}^{\infty} K_{i+j}(\hat{\Omega} - \Omega)K'_i$ also converges to zero according to the ergodicity assumption. The rate of convergence might be slower which, however, does not conflict with the rate of convergence of the subspace estimators as follows from the following result which is proved in the appendix.

Theorem 1 *Let $(\tilde{y}_t)_{t \in \mathbb{Z}}$ be generated by a stable and strictly minimum-phase state space system (A_0, K_0, C_0) of order n , where the innovations process $(\varepsilon_t)_{t \in \mathbb{Z}}$ fulfills assumptions 1. Assume that \tilde{y}_t is not directly observed, but only $y_t = \tilde{y}_t + DT_t$ is observed, where $(T_t)_{t \in \mathbb{Z}}$ denotes an observed process for which assumptions 3 hold. Let \hat{y}_t denote the residuals of y_t regressed onto T_t . Let $\theta_{0,i} = \varphi_i^{-1}(\pi(A_0, K_0, C_0))$ denote the true parameter vector where the index i is chosen such that $\pi(A_0, K_0, C_0)$ is an interior point of $M(n; i)$. Further let $\hat{\theta}_i = \varphi_i^{-1}(\pi(\hat{A}, \hat{K}, \hat{C}))$, where $(\hat{A}, \hat{K}, \hat{C})$ denotes the subspace estimator based on $\hat{y}_t, t = 1, \dots, T$ described above using the true order, which depends on the choice of \hat{W} and f, p . Then under assumption 2 $\hat{\theta}_i$ is well defined (a.s. for large T) for $\hat{W} = (\hat{W}_f^+)' \hat{W}_f^+$. Further $\hat{\theta}_i \rightarrow \theta_{0,i}$ a.s. If $p \geq -d \log T / (2 \log \rho_0)$, $d > 1$ arbitrary, where $\rho_0 = |\lambda_{\max}(A_0 - K_0 C_0)|$, then $\|\hat{\theta}_i - \theta_{0,i}\| = O(Q_T)$.*

The theorem gives a close to maximal result in terms of the speed of convergence of the estimator being of the form of an LIL except for the fact that the constant in the $O(\cdot)$ statement is not evaluated. The same rate applies also in the conditionally homoskedastic case, hence the algorithms are robust with respect to the conditional heteroskedasticity of the innovations subject to assumptions 1.

Note that the result holds true for any $k(z) \in M(n)$. Only the coordinate system used to present the result changes for different pieces $M(n; i)$ of $M(n)$. Also the presentation in terms of the state space form is not essential. The results in Chapter 2 of Hannan and Deistler (1988) show that it is simple to obtain the corresponding results of the ARMA representations for the echelon forms used in this paper. Therefore the whole paper could be written avoiding the mentioning of the state space representation at the expense of making the presentation of the subspace algorithms complicated.

Beside consistency also the asymptotic distribution is of interest. As in the conditionally homoskedastic case the estimators can be shown to follow a CLT. Without additional assumptions on the behaviour of the conditional moments it is not possible to obtain simple expressions for the asymptotic variance.

Theorem 2 *Let the assumptions of Theorem 1 hold. Assume that $p \geq -d \log T / (2 \log \rho_0)$ for some $d > 1$ holds a.s. Then under assumptions 2 and 3, $\sqrt{T} \left(\hat{\theta}_i - \theta_{0,i} \right) \xrightarrow{d} Z$ where Z is multivariate normally distributed with mean zero and some variance V .*

This theorem highlights the major difference between the conditionally homoskedastic and the conditionally heteroskedastic case: The asymptotic distribution remains to be Gaussian, but the variance is hard to quantify (although straightforward to estimate). This is analogous to the prediction error case and pseudo ML estimation as discussed in (Hannan and Deistler, 1988, section 4.3).

For the conditionally homoskedastic case the equivalence to prediction error minimization has been proved in Bauer (2000). It turns out that the equivalence also prevails for the conditionally heteroskedastic situation.

Theorem 3 *Let the assumptions of Theorem 1 hold. Furthermore let the components of T_t be harmonic processes, i.e. $T_{t,i} = \mathcal{R}(\mu_i^t T_{0,i}^c)$ for some $|\mu_i| = 1, \mu_i \in \mathbb{C}, T_{0,i}^c \in \mathbb{C}$. Assume that the CCA algorithm using $f = p, p \geq -d \log T / (2 \log \rho_0), d > 1$ arbitrary, $p = O((\log T)^a)$ for some $a < \infty$ is used to obtain the estimate $\hat{\theta}_i$. Let θ_i denote the estimate obtained by minimizing $L(\hat{y}_t, \theta, \hat{\omega})$. Then $\sqrt{T} \|\hat{\theta}_i - \theta_i\| = o_P(1)$.*

All prior results deal with the case that the true order of the system is known. In practice, however, the order typically is estimated. For subspace methods a number of order estimation techniques have been developed, which use the information contained in the singular values. The main tool in the derivation of consistency of the order estimators again was the uniform convergence of the sample covariance sequence. Hence also these results can be generalized.

Theorem 4 *Let the assumptions of Theorem 1 hold. Assume that the order is estimated using $IC(n)$ for $\|\hat{R}_n\|^2 = \hat{\sigma}_{n+1}^2$. Let \hat{n} be obtained as the minimal minimizing argument of $IC(n)$ over $0 \leq n \leq \min(fs, ps)$. Assume further that assumptions 2 and 3 hold. Then $C_T / (fp \log \log T) \rightarrow \infty$ is a sufficient condition for consistency of \hat{n} , i.e. $\hat{n} \rightarrow n_0$ a.s. where n_0 denotes the true order of the data generating process.*

Again the proof of this theorem is given in the appendix.

Summing up the results presented so far essentially the robustness of subspace methods with respect to the conditional heteroskedasticity has been showed. The estimators remain consistent, asymptotically normal (though the asymptotic variance depends on the nature of the conditional heteroskedasticity) and in the CCA case also asymptotically equivalent to prediction error methods when allowing for some conditional heteroskedasticity. It has also been showed that the CCA procedure does not achieve optimal accuracy. Hence the value added from this procedure might be doubted. From my point of view the value added lies in the relatively low cost for the estimation of the model of the mean, which at the very least provides an initial guess for a subsequent pseudo likelihood analysis. Since the procedure provides consistent estimates of the innovations these can be used in order to specify the model for the conditional variance and also to obtain a consistent estimate thereof for initialization of the full likelihood estimation. This is analogous to the discussion in section 4 of Mills (1994) for the univariate case. Especially in a multivariate framework this is seen to be a useful tool.

The subspace estimators of the innovations sequence can be useful in a number of ways: First of all, the estimated innovations sequence can be used on a purely descriptive basis in order to derive intuition about the nature of time variability of the conditional variance. Secondly one could fix the model for the mean using the CCA estimates of the innovations, $\hat{\varepsilon}_t$ say, and specify and estimate the model for the variance based on the minimization of the criterion

function

$$L(\hat{y}_t, \hat{\theta}, \omega) = \frac{1}{T} \sum_{t=1}^T \left(\log \det(H_t(\omega, \hat{\theta})) + \hat{\varepsilon}_t(\hat{\theta})' H_t(\omega, \hat{\theta})^{-1} \hat{\varepsilon}_t(\hat{\theta}) \right) \quad (5)$$

where $H_t(\omega, \hat{\theta})$ denotes the conditional variance based on the parameter vector $\omega \in S$ and the innovation estimates $\hat{\varepsilon}_t(\hat{\theta})$ obtained using the Kalman filter based on the parameter estimates $\hat{\theta}$ and zero initial conditions. Under the assumption of $\sup_{t \in \mathbb{N}} \sup_{\omega \in S, \theta \in T_i} \|H_t(\omega, \theta)^{-1}\| < C < \infty$ a.s. and $\sup_{\omega \in S} T^{-1} \sum_{t=1}^T \|H_t(\omega, \hat{\theta}) - H_t(\omega, \theta_0)\|_{Fr}^2 \rightarrow 0$ where $\|\cdot\|_{Fr}$ denotes the Frobenius norm it can be showed that the difference between $L(\hat{y}_t, \hat{\theta}, \omega)$ and $L(\hat{y}_t, \theta_{0,i}, \omega)$, which uses the true innovations ε_t in place of the estimated innovations $\hat{\varepsilon}_t(\hat{\theta})$ converges to zero uniformly in $\omega \in S$. Therefore typical consistency proofs for the estimation of ω based on $L(\hat{y}_t, \theta_{0,i}, \omega)$ also apply for $L(\hat{y}_t, \hat{\theta}, \omega)$ under suitable assumptions on the model for the conditional variance ensuring the above mentioned conditions. The asymptotic distribution of the estimator $\hat{\omega}$ however depends on whether $\theta_{0,i}$ or $\hat{\theta}$ is used in the estimation (cf. Weiss, 1986, for the univariate case). For a given variance model it seems to be possible in principle to find the asymptotic distribution adjusted for the estimation of the innovations. Here we will describe a different approach.

As a demonstration of this approach we derive tests for the structure of the covariance model using the estimated innovations in an ARCH(p) framework. Rather than finding the distribution of tests based on adjusted variances we provide estimators whose asymptotic distribution is invariant to the pre-estimation of the innovations:

Theorem 5 *Let the assumptions of Theorem 2 hold. Let $(\varepsilon_t)_{t \in \mathbb{Z}}$ denote the true innovation process and $(\hat{\varepsilon}_t)_{t \in \mathbb{Z}}$ the subspace estimators thereof obtained under the conditions of Theorem 2, i.e. the estimators of the innovations corresponding to the Kalman filter according to the estimate $\hat{\theta}_i$ using zero initial conditions. Consider the regression model (vech denotes the operator of stacking the lower triangular part of the matrix into a vector)*

$$\text{vech}[\varepsilon_t \varepsilon_t'] = C + \sum_{j=1}^p \alpha_j \text{vech}[\varepsilon_{t-j} \varepsilon_{t-j}'] + u_t = [C, \alpha_1, \dots, \alpha_p] x_t + u_t$$

in a multivariate ARCH(p) model such that $\mathbb{E}\{u_t | \mathcal{F}_{t-1}\} = 0$. Let $\beta = [C, \alpha_1, \dots, \alpha_p] \in \mathbb{R}^{s(s+1)/2 \times (1+s(s+1)p/2)}$ denote the parameter matrix. The equation above defines x_t the vector of regressors. Let $z_t \in \mathbb{R}^z, z \geq [s(s+1)p/2 + 1]$ denote instrumental variables, which are assumed to be \mathcal{F}_{t-p-1} measurable, strictly stationary, and ergodic processes such that $\mathbb{E}x_t z_t'$ and $\mathbb{E}z_t z_t'$ are of full row rank. Furthermore it is assumed that $\sup_t \|z_t\| < M < \infty$ a.s. for some constant M . Assume that the vector process $[\varepsilon_t', T_t', z_t']'$ is ergodic and strictly stationary. Let $\hat{\beta}_{IV}$ denote the corresponding IV estimate defined as

$$\hat{\beta}_{IV} = \left\{ \langle \text{vech}[\varepsilon_t \varepsilon_t'], z_t \rangle \langle z_t, z_t \rangle^{-1} \langle z_t, x_t \rangle (\langle x_t, z_t \rangle \langle z_t, z_t \rangle^{-1} \langle z_t, x_t \rangle)^{-1} \right\}$$

Further consider the analogous equation

$$\text{vech}[\hat{\varepsilon}_t \hat{\varepsilon}_t'] = C + \sum_{j=1}^p \alpha_j \text{vech}[\hat{\varepsilon}_{t-j} \hat{\varepsilon}_{t-j}'] + \hat{u}_t.$$

Let $\tilde{\beta}_{IV}$ denote the corresponding IV estimate using identical instruments z_t . Then $\sqrt{T}(\hat{\beta}_{IV} - \tilde{\beta}_{IV}) \xrightarrow{P} 0$ and therefore tests based on the asymptotic distribution of the

estimated parameter vector remain asymptotically valid if calculated using the estimated innovations $\hat{\varepsilon}_t$ in place of the true innovations ε_t . Let the asymptotic variance matrix of $\text{vec}(\hat{\beta}_{IV})$ be estimated as

$$\left[\tilde{\Sigma}_{XZ} \otimes I \right] \left[\frac{1}{T} \sum_{t=p+1}^T (z_t z_t' \otimes \tilde{u}_t \tilde{u}_t') \right] \left[\tilde{\Sigma}'_{XZ} \otimes I \right]$$

where $\tilde{\Sigma}_{XZ} = (\langle \hat{x}_t, z_t \rangle \langle z_t, z_t \rangle^{-1} \langle z_t, \hat{x}_t \rangle)^{-1} \langle \hat{x}_t, z_t \rangle \langle z_t, z_t \rangle^{-1}$, \hat{x}_t denotes the estimates of x_t based on $\hat{\varepsilon}_t$ and $\tilde{u}_t = \tilde{E}_t - \tilde{\beta}_{IV} \hat{x}_t$. This estimator consistently estimates the variance of the limiting normal distribution for $\sqrt{T}(\hat{\beta}_{IV} - \beta)$.

This result contrasts the theory on post-estimation testing in Weiss (1986) for the univariate ARMA-GARCH case. Rather than deriving the adjusted distribution of the test statistic we show that the asymptotic distribution of the instrumental variables estimator based on sufficiently lagged estimates is robust with respect to the estimation of the innovations sequence. As a consequence standard software can be used to conduct the proposed test which is seen to be an advantage.

Note that the weighting by $\langle z_t, z_t \rangle^{-1}$ corresponds to the optimal weight in the conditionally homoskedastic situation which need not necessarily be a good choice in the conditionally heteroskedastic case. A more general weighting could be included in the result. Since this result is only seen as a prototypical result for demonstration purposes we refrain from elaborating on this.

As for all instrumental variables estimators the question of suitable instruments arises. A second concern is the loss of accuracy due to the usage of suboptimal estimators. In the present case an attractive set of instruments is given by a subset of the variables ⁶ $\text{vec}[\hat{y}_{t-j} \hat{y}'_{t-i}]$, $i, j > p$. This choice of instruments is attractive since typically in financial time series the correlation in the returns dies out very fast (in daily return series typically the evidence for correlations is rather weak, in high frequency data correlations are restricted to few periods, say up to half a day), whereas the correlations of the squared returns die out slowly. Typically values of $\alpha + \beta$ close to one for GARCH(1,1) models have been observed. Therefore it seems likely that ε_t can be recovered using only a few past lags of \hat{y}_t , whereas the correlation of $\varepsilon_t \varepsilon_t'$ with $\varepsilon_{t-p} \varepsilon_{t-p}'$ is still reasonably high. The measurability assumption is guaranteed for sufficiently lagged \hat{y}_t 's. The remaining assumptions on the instrumental variables of course are model dependent and cannot be verified in general.

5 Simulations

In this section three simulation experiments are presented dealing with two major points: The first point refers to the computational comparison: The simulations below show that subspace algorithms are computationally much cheaper than prediction error methods while providing estimates of roughly the same accuracy. The general setup is to either increase the sample size while keeping the structure (dimension of the output process and the order of the state space system) fixed or to keep the sample size fixed and let the number of outputs grow. Note that the size of the data sets considered here is smaller than typical applications would require. This is done in order to keep the computational burden for the prediction error methods in an acceptable range. The second point we are trying to make is that the

⁶In order to comply with the boundedness assumption a trimming might be necessary.

subspace estimates of the innovations can be used for purposes of specifying the model for the conditional variances.

The setup for the first two experiments is as follows: For each experimental condition a number of time series of length $T + 100$ are created and the first 100 observations are omitted in order to decrease the dependence on initial effects. One state space system of output dimension s and state dimension n is generated randomly, where the strict minimum-phase condition is imposed and $|\lambda_{max}(A)| < 0.3$ is imposed in order to match typical high frequency data characteristics. All eigenvalues of A are chosen to be real, which might be seen as a limitation to randomness. The model for the conditional heteroskedasticity is of the constant correlation GARCH type with random nonnegative coefficients. The coefficients for the univariate GARCH models for the diagonal elements obey $\alpha + \beta < 1/2$. For each experiment only one system has been generated in order to make comparison across different conditions possible. In the first experiment the sample size T has been chosen to be 1000, 2500, 5000 and 10000 respectively, $s = 6$ and $n = 2$ is used. In the second experiment $T = 1000$ is fixed and $s = 10, 20, 30, 40$ is used and $n = 6$. In both experiments the accuracy is measured in terms of the sum of the two norm of the estimation errors in the estimated impulse response coefficients $\hat{C}\hat{A}^{j-1}\hat{K}, j = 1, 2, \dots, 2n - 1$. For each time series two estimates are calculated using MATLAB: The prediction error estimate is obtained using the command `mod = pem(iddata(y(101:end,:),n),n)`; contained in the system identification toolbox of MATLAB. This algorithm obtains an initial guess using a subspace algorithm called N4SID (for a discussion see Van Overschee and DeMoor, 1994) to start a Gauss-Newton search for the optimum. Therefore it comes at no surprise that the numerical load is bigger for `pem`. The CCA estimate is obtained using a self written program. No preprocessing of the data such as mean extraction is performed. The consequences of these choices are the following: We compare a commercial product (`pem`) with a code written for academic purposes which does not include the usual consistency checks, error handling code etc. In particular it does also not contain the calculation of the estimate of the covariance matrix. Therefore the computational burden for CCA is lower than the burden for `pem` in this respect. This does not make up for a significant part, however. The second implication is that the comparison really is based on the numerical procedures, not on the theoretical concepts. In particular there is no guarantee that `pem` delivers the correct minimizing argument of the criterion function. In fact it happened in some (undocumented) cases that `pem` seems to be caught in a local minimum resulting in overly large errors. Another point is in order here: Since a particular algorithm is used in the comparison the simulations are of course subject to the critique that a different procedure might not suffer from the problems encountered in `pem`. To a certain extent such a critique is of course adequate, especially procedures based on maximizing the Whittle likelihood or similar procedures starting from the estimated covariance sequence will probably show better performance in terms of computations. Numerical issues with respect to illconditioning of the Hessian for high dimensional parameter sets remain valid, however.

The results of the first experiment are given in Table 1. In this experiment `pem` performs better than CCA for the smaller sample sizes while for the large sample sizes the difference in estimation error is negligible as predicted by theory. For both procedures consistency can be observed. In all cases CCA outperforms `pem` in terms of the number of computations where the difference is significant. Somewhat surprisingly the difference decreases with increasing sample size. Undocumented additional simulations show that for CCA there is a trade-off between the computational time and the achieved accuracy related to the allowed upper bound for the autoregression in order to obtain the estimate \hat{p}_{AIC} . In the documented results this upper

| | Mean error | | | Comp. time (in sec.) | | |
|-------------|------------|--------|-------|----------------------|---------|-------|
| | CCA | pem | Ratio | CCA | pem | Ratio |
| $T = 1000$ | 0.7085 | 0.4440 | 1.60 | 0.1146 | 1.4287 | 12.47 |
| $T = 2500$ | 0.3057 | 0.2887 | 1.06 | 0.4040 | 3.0433 | 7.53 |
| $T = 5000$ | 0.2063 | 0.2030 | 1.02 | 1.1437 | 5.5747 | 4.87 |
| $T = 10000$ | 0.1346 | 0.1346 | 1.00 | 3.7279 | 10.2962 | 2.76 |

Table 1: Experiment 1: Comparison of CCA to pem in MATLAB system identification toolbox. $s = 6, n = 2$, numbers are based on 1000 simulations for each entry. Computation times are given for estimation on one data set.

bound is chosen to be $\sqrt{T}/2$ so that potentially covariance estimates up to lag $2\sqrt{T}$ are used in the subspace estimation. Almost all the computations for the two larger sample sizes are spent on obtaining estimates for the covariance sequence (94% for $T = 5000$ and more than 97 % for $T = 10000$). Choosing the upper bound differently (such as e.g. $4 \log T$) reduces the computational load at large sample sizes, however, potentially limits the approximation accuracy leading to a computations-accuracy tradeoff in this example. Note, however, that in all cases the computation times for pem are acceptable. The conclusions drawn from the first experiment (and undocumented others) is that in data sets of the given dimensions pem still is feasible computationally and results on average in better accuracy of the estimates. In these situations CCA might serve as an initial guess but prediction error methods appear to be superior. However, bear in mind that in the conditionally heteroskedastic case the model for the mean is estimated in any case only as an initial guess for subsequent pseudo likelihood maximization including the model for the conditional variance.

The second experiment used the same setup as the first experiment except that in this experiment the output dimension is increased for fixed sample size $T = 1000$ and order $n = 6$. For $s = 10$ and $s = 20$ we generate 100 time series and perform the estimation, whereas the number of simulation experiments are limited to 50 for $s = 30$ and 25 for $s = 40$. The results are documented in Table 2. Here we report only the trimmed mean estimation error rather than the mean error, since in a number of cases pem resulted in excessively large errors, which are seen to be due to problems of finding the global minimum rather than problems for the prediction error estimates themselves. Initialization in pem is done using a different subspace algorithm, which seems to provide sometimes unreliable estimates. CCA did not show the same problems in our simulations ⁷. In these examples the estimation accuracy is (statistically) indistinguishable for $s = 10$ and $s = 20$ for the trimmed means while for $s = 30$ and $s = 40$ CCA clearly is outperformed by pem. The computation times, however, differ drastically. pem is heavily affected by the increase of the parameter dimension from 24 parameters in the first experiment to 120 for $s = 10$ and 240 for $s = 20$. The computational comparison now is strongly favoring CCA even for the smaller output dimensions and is becoming drastic for the two large output sizes. For $s = 20$ pem needs more than a minute. For $s = 40$ executing pem takes more than 15 minutes even without order estimation, which typically requires a number of models to be estimated. This seems to be prohibitive. CCA on the other hand still uses only a few seconds ⁸.

⁷Of course, CCA could have been used as initial guess to pem. But the main point in this comparison is the computational load which does not rely heavily on the initial estimate in this experiment.

⁸Adding order estimation using $IC(n)$ introduces a negligible additional load.

| | Trimmed mean error | | | Comp. time (in sec.) | | |
|---------------------|--------------------|--------|-------|----------------------|--------|--------|
| | CCA | pem | Ratio | CCA | pem | Ratio |
| $s = 10$ (100 rep.) | 1.8399 | 1.9008 | 0.97 | 0.22 | 35.78 | 162.63 |
| $s = 20$ (100 rep.) | 0.9638 | 0.9483 | 1.02 | 0.46 | 77.64 | 168.78 |
| $s = 30$ (50 rep.) | 1.09 | 0.99 | 1.11 | 0.87 | 288.20 | 332.63 |
| $s = 40$ (25 rep.) | 3.25 | 2.90 | 1.12 | 1.41 | 854.09 | 605.73 |

Table 2: Experiment 2: Comparison of CCA to pem in MATLAB system identification toolbox. $n = 6$ and $T = 1000$. Trimming omits 5% of the observations. Computation times are given for estimation on one data set.

| Test based on $\varepsilon_t/\hat{\varepsilon}_t$ | Size | | | | Power | | | |
|--|---------|---------|---------|---------|---------|---------|---------|---------|
| | acc/acc | acc/rej | rej/acc | rej/rej | acc/acc | acc/rej | rej/acc | rej/rej |
| $T = 1000$ | 92.0 | 1.4 | 1.3 | 5.3 | 57.1 | 4.3 | 3.2 | 35.4 |
| $T = 5000$ | 93.2 | 1.2 | 0.7 | 4.9 | 39.5 | 1.5 | 1.1 | 57.9 |
| $T = 10000$ | 93.8 | 0.9 | 0.4 | 4.9 | 17.9 | 1.4 | 0.6 | 80.1 |

Table 3: Experiment 3: Performance of the testing procedures. All numbers are percentages on the basis of 1000 simulations with $s = n = 2$ of a BEKK-ARCH model using $n = m = 1$. Nominal size 95%. Acceptance (acc) and rejection (rej) rates for the tests based on the true innovations ε_t and estimated innovations $\hat{\varepsilon}_t$.

Finally also the inference based on the estimated innovations is investigated. In this respect a bivariate BEKK-ARCH model is used in the simulations. $n = m = 1$ is used in the specification of the variance model. In the notation of Theorem 5 we thus have $p = 1$. The model for the mean is drawn as before randomly. Two different tests will be considered: In both cases $p = 2$ is used in the estimation with no restrictions to the ARCH model. The first test considers testing the null of $p = 1$ against the alternative of $p = 2$, hence exploring the size properties of the tests. The second situation tests the null of $p = 0$ (conditional homoskedasticity) against the alternative of $p = 2$ exploring the power of the tests. In all cases 1000 time series are simulated. $s = n = 2$ is used and T is varied from $T = 1000, T = 5000$ to $T = 10000$. The results can be seen in Table 3. The results of this experiment are both affirmative and to a certain extent discouraging: The procedure using the estimated innovation sequence $(\hat{\varepsilon}_t)_{t=1,\dots,T}$ shows very similar behaviour to the tests based on $(\varepsilon_t)_{t=1,\dots,T}$. In all simulations the two tests gave the same result in more than 92% of the cases and for $T = 10000$ even in 98% of the cases. Due to the high noise level in these models (due to using $\varepsilon_t \varepsilon_t'$ in place of the conditional variance) very large sample sizes are required in order for the tests to have good size and power properties. Even at $T = 10000$ only in slightly more than 80% of the cases evidence for heteroskedasticity is found. Therefore our conclusions from these experiments is that the asymptotical equivalence of the tests is reflected in the test results already in medium sized samples, however, this is contrasted by the fact that the test performance in absolute terms is rather poor in this setting.

6 Conclusions

In this paper the asymptotic properties of subspace algorithms have been extended to the conditionally heteroskedastic case by a more careful study of the assumptions sufficient for the previously published results. The main message is that – except for the expressions for the asymptotic variance – all results achieved for the conditionally homoskedastic case carry over also to the case of conditionally heteroskedastic innovations. In particular the equivalence of CCA estimators to prediction error estimators prevails. This implies that the CCA estimators are not asymptotically efficient in this case. Hence they should not be used as final estimators. Nevertheless they are seen as valuable tools, as they provide estimates for multivariable systems in situations, where prediction error minimization is numerically not feasible as documented in the simulations section. This can be used in order to obtain estimates of the residuals on which a subsequent modeling of the conditional variance can be based.

Acknowledgements

Financial support via the Austrian FWF under project number P-14438INF and the Max Kade foundation is gratefully acknowledged. Part of this work has been done while the author was holding a post-doc position at the Cowles Foundation for Research in Economics, Yale University, New Haven, CT, USA, whose hospitality is appreciated. Finally I would like to thank Wolfgang Scherrer, Martin Wagner and Michael McAleer for fruitful discussions on the topics of this paper.

References

- Baillie, R. and T. Bollerslev (1990). Intra-day and inter-market volatility in foreign exchange rates. *Review of Economic Studies* **58**, 565–585.
- Bauer, D. (1998). Some Asymptotic Theory for the Estimation of Linear Systems Using Maximum Likelihood Methods or Subspace Algorithms. PhD thesis. TU Wien, Austria.
- Bauer, D. (2000). Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs. Technical report. Institute f. Econometrics, Operations Research and System Theory, TU Wien. Submitted to Journal of Time Series Analysis, available at <http://www.eos.tuwien.ac.at/Oeko/DBauer/jtsa.pdf>.
- Bauer, D. (2001). Order estimation for subspace methods. *Automatica* **37**, 1561–1573.
- Bauer, D. (2003). Subspace methods. In: *Proceedings of the IFAC Symposium on System Identification, SYSID'03*. Rotterdam, The Netherlands.
- Bauer, D. and L. Ljung (2002). Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms. *Automatica* **38**, 763–773.
- Bauer, D., M. Deistler and W. Scherrer (1999). Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs. *Automatica* **35**, 1243–1254.

- Bauer, D., M. Deistler and W. Scherrer (2000). On the impact of weighting matrices in subspace algorithms. In: *Proceedings of the IFAC Conference SYSID'00*. Santa Barbara, California.
- Berndt, E., B. Hall, R. Hall and J. Hausman (1974). Estimation inference in nonlinear structural models. *Annals of Economic and Social Measurement* **4**, 653–665.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.
- Bougerol, P. and N. Picard (1992). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* **52**, 115–127.
- Boussama, F. (1998). Ergodicity, Mixing and Estimation in GARCH Models. PhD thesis. University of Paris 7.
- Camba-Mendez, G. and G. Kapetanios (2001). Testing the rank of the Hankel covariance matrix: A statistical approach. *IEEE Transactions on Automatic Control* **46**, 331–336.
- Comte, F. and O. Lieberman (2003). Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis* **84**, 61–84.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.
- Deistler, M., K. Peternell and W. Scherrer (1995). Consistency and relative efficiency of subspace methods. *Automatica* **31**, 1865–1875.
- Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.
- Engle, R. and K. Kroner (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory* **11**, 122–150.
- Findley, D., B. Pötscher and C. Wei (2001). Uniform convergence of sample second moments of families of time series arrays. *The Annals of Statistics* **29**, 815–838.
- Gourieroux, Chr. (1997). *ARCH Models and Financial Applications*. Springer Series in Statistics.
- Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. John Wiley. New York.
- Jeantheau, T. (1998). Strong consistency of estimators for multivariate ARCH models. *Econometric Theory* **14**, 70–86.
- Larimore, W. E. (1983). System identification, reduced order filters and modeling via canonical variate analysis. In: *Proc. 1983 Amer. Control Conference 2*. (H. S. Rao and P. Dorato, Eds.). Piscataway, NJ. pp. 445–451.
- Ling, S. and M. McAleer (2002). Necessary and sufficient moment conditions for the GARCH(r,s) and asymmetric power GARCH(r,s) models. *Econometric Theory* **18**, 722–729.

- Ling, S. and M. McAleer (2003). Asymptotic theory for a vector ARMA-GARCH model. *Econometric Theory* **19**, 278–308.
- Mills, T. (1994). *The Econometric Modeling of Financial Time Series*. Cambridge University Press.
- Nelson, D. (1991). Conditional heteroskedasticity in asset returns. A new approach. *Econometrica* **59**, 347–370.
- Peternell, K. (1995). Identification of Linear Dynamic Systems by Subspace and Realization-Based Algorithms. PhD thesis. TU Wien.
- Rahbek, A., E. Hansen and J. Dennis (2003). ARCH innovations and their impact on cointegration rank testing. Technical report. Department of Theoretical Statistics, University of Copenhagen.
- Van Overschee, P. and B. DeMoor (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica* **30**, 75–93.
- Verhaegen, M. (1994). Identification of the deterministic part of mimo state space models given in innovations form from input-output data. *Automatica* **30**(1), 61–74.
- Weiss, A. A. (1986). Asymptotic theory for ARCH models: Estimation and testing. *Econometric Theory* **2**, 107–131.

A Proofs

Throughout the proofs estimators of a quantity x will be denoted using \hat{x} . For two sequences $(a_t)_{t \in \mathbb{Z}}, (b_t)_{t \in \mathbb{Z}}$ we will use the notation $\langle a_t, b_t \rangle = \frac{1}{T} \sum_{t=1}^T a_t b_t'$ where it is understood that unavailable observations are replaced by zeros. An observation is unavailable if it is built using y_t for $t < 1$ or $t > T$. Expressions where the noise variance matrix Ω is replaced by $\langle \varepsilon_t, \varepsilon_t \rangle$ are denoted using a superscripted \cdot . Recall that $\tilde{y}_t = y_t - DT_t$ and $\hat{y}_t = y_t - \langle y_t, T_t \rangle \langle T_t, T_t \rangle^{-1} T_t$. Recall that $F_T = O(g_T)$ means $\limsup_{T \rightarrow \infty} \max_{i,j} |F_{T,i,j}|/g_T \leq C$ a.s. for some constant $C < \infty$. Here $F_{T,i,j}$ denotes the (i, j) entry of the matrix F_T . This notation will be used also for matrices of dimensions increasing with sample size T . Note that this definition differs from conventional usage, where the norm of the matrix is bounded rather than the maximal entry. Throughout the proof we will often use the following result:

Lemma 1 *Let $\hat{\Phi} \in \mathbb{R}^{a \times b}$ and $\hat{\Psi} \in \mathbb{R}^{b \times c}$ where the dimensions a, b, c possibly depend on the sample size. Assume that there exist matrices $\Phi \in \mathbb{R}^{a \times b}$ and $\Psi \in \mathbb{R}^{b \times c}$ such that $\hat{\Phi} - \Phi = O(a_T)$ and $\hat{\Psi} - \Psi = O(b_T)$. Moreover assume that there exists a constant $C < \infty$ such that $\max(\limsup_{a,b} \|\Phi\|_\infty, \limsup_{b,c} \|\Psi'\|_\infty) < C$. Then $\hat{\Phi}\hat{\Psi} - \Phi\Psi = O(\max(a_T, b_T, a_T b_T b))$. If Φ is equal to zero, then $\hat{\Phi}\hat{\Psi} = O(\max(a_T, a_T b_T b))$.*

Proof: Note that $\hat{\Phi}\hat{\Psi} - \Phi\Psi = (\hat{\Phi} - \Phi)\hat{\Psi} + \Phi(\hat{\Psi} - \Psi) = (\hat{\Phi} - \Phi)(\hat{\Psi} - \Psi) + (\hat{\Phi} - \Phi)\Psi + \Phi(\hat{\Psi} - \Psi)$. Due to the uniform bound on the infinity norms the latter two terms are of the order $O(a_T)$ and $O(b_T)$ respectively. The entries in the first term are of order $O(a_T b_T b)$ being the sum of b products of elements in $\hat{\Phi} - \Phi$ and $\hat{\Psi} - \Psi$ respectively. If Φ is zero then the last term is zero and hence b_T does not appear. \square

Note that extensions to products containing more than two matrices are straightforward:

$\hat{\Phi}\hat{\Psi}\hat{\Theta} - \Phi\Psi\Theta = (\hat{\Phi}\hat{\Psi} - \Phi\Psi)\hat{\Theta} + \Phi\Psi(\hat{\Theta} - \Theta) = (\hat{\Phi}\hat{\Psi} - \Phi\Psi)\Theta + (\hat{\Phi}\hat{\Psi} - \Phi\Psi)(\hat{\Theta} - \Theta) + \Phi\Psi(\hat{\Theta} - \Theta) = O(\max(a_T, b_T, c_T))$ under the assumption that the order of convergence of each of the matrices is faster than the increase in the dimensions, i.e. e.g. $b_T b \rightarrow 0$ holds, as will always be the case for the expressions below. Also suitable assumptions on the matrices must ensure that the infinity norm of all occurring matrices remains bounded, e.g. in the example given above $\|\Phi\Psi\|_\infty$.

A.1 Proof of Theorem 1

The proof parallels the discussion of (Bauer, 2000, Lemma A.1) but replaces the true covariance sequence with the sequence $\dot{\gamma}_j$. The main facts used in the proof will be provided in lemmas.

The central estimate in subspace procedures is the estimate $\hat{\beta}$ of $\mathcal{O}_f\mathcal{K}_p$. Its properties in the various estimation conditions are stated in the first lemma:

Lemma 2 *Let the assumptions of Theorem 1 hold. Then $\|\dot{\Gamma}_\infty^-\|_\infty < \infty$, $\sup_{1 \leq p \leq ((\log T)^a)} \|\dot{\Gamma}_p^-\|_2 < M$, $\sup_{1 \leq p \leq ((\log T)^a)} \|(\dot{\Gamma}_p^-)^{-1}\|_2 < M$ a.s. and $\sup_{1 \leq p \leq ((\log T)^a)} \|(\dot{\Gamma}_p^-)^{-1}\|_\infty < M$ a.s. for some constant $M < \infty$ and arbitrary $a < \infty$. The same bounds hold a.s. for large enough T for $\dot{\Gamma}_p^-$ replaced with $\langle Y_{t,p}^-, Y_{t,p}^- \rangle$.*

Furthermore $\hat{\beta} - \mathcal{O}_f\mathcal{K}_p = O(\max(Q_T, \rho^p))$ for arbitrary $\rho_0 < \rho < 1$ for $\rho_0 = |\lambda_{\max}(A - KC)|$ where (A, K, C) denotes a representation of the true system. Hence for $p \geq -d \log T / (2 \log \rho_0)$, $d > 1$ it holds that $\hat{\beta} - \mathcal{O}_f\mathcal{K}_p = O(Q_T)$.

Proof: The norm bounds can be found in Theorem 6.6.10 and 6.6.11. of Hannan and Deistler (1988).

The proof for the order of convergence of $\hat{\beta} - \mathcal{O}_f\mathcal{K}_p$ will be given for the case $\hat{D} = D$ first. In the following we will use the symbols $Y_{t,f}^+$ and $Y_{t,p}^-$ neglecting the fact that these are built using \tilde{y}_t in place of \hat{y}_t . This should not cause any confusion. Let $\dot{\beta} = \dot{\mathcal{H}}_{f,p}(\dot{\Gamma}_p^-)^{-1}$ where $\dot{\mathcal{H}}_{f,p} = [\dot{\gamma}_{i+j-1}]_{i=1,\dots,f,j=1,\dots,p}$ is the covariance of $Y_{t,f}^+$ and $Y_{t,p}^-$ with Ω replaced by $\dot{\Omega}$ such that

$$\hat{\beta} - \dot{\beta} = \langle Y_{t,f}^+, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} - \dot{\mathcal{H}}_{f,p}(\dot{\Gamma}_p^-)^{-1}.$$

Due to the uniform convergence of the estimated covariance sequence, see (2), it follows that $\langle Y_{t,f}^+, Y_{t,p}^- \rangle - \dot{\mathcal{H}}_{f,p} = O(Q_T)$ and $\langle Y_{t,p}^-, Y_{t,p}^- \rangle - \dot{\Gamma}_p^- = O(Q_T)$. Using Lemma 1 the invertibility of $\dot{\Gamma}_p^-$ a.s. for large T together with boundedness (uniformly in p) of $\|(\dot{\Gamma}_p^-)^{-1}\|_\infty$ imply that $\hat{\beta} - \dot{\beta} = O(Q_T)$. Furthermore $\mathcal{O}_f\mathcal{K}_p - \dot{\beta} = O(\rho^p)$ for arbitrary $\rho_0 < \rho < 1$ follows from

$$[\dot{\mathcal{H}}_{f,p}(\dot{\Gamma}_p^-)^{-1}, 0](\dot{\Gamma}_\infty^-)_p - \dot{\mathcal{H}}_{f,\infty}(\dot{\Gamma}_\infty^-)^{-1}(\dot{\Gamma}_\infty^-)_p = 0$$

a.s. where $(\dot{\Gamma}_\infty^-)_p$ denotes the first p block columns of $\dot{\Gamma}_\infty^-$ (cf. also Bauer *et al.*, 1999, Lemma 6, where the same result is showed using the true noise covariance rather than $\dot{\Omega}$). This equation shows that $\hat{\beta} - \mathcal{O}_f\mathcal{K}_p = \mathcal{O}_f(A - KC)^p \mathcal{K}_\infty (\dot{\Gamma}_\infty^-)_{2,p} (\dot{\Gamma}_p^-)^{-1}$ using $\dot{\mathcal{H}}_{f,\infty}(\dot{\Gamma}_\infty^-)^{-1} = \mathcal{O}_f\mathcal{K}_\infty$ where $(\dot{\Gamma}_\infty^-)_{2,p}$ denotes the matrix obtained from $(\dot{\Gamma}_\infty^-)_p$ by omitting the first p block rows. The order of convergence then follows from the uniform bound (in p) on the smallest eigenvalue of $\dot{\Gamma}_p^-$, $\|\dot{\Gamma}_\infty^-\|_\infty < \infty$ and $\|(\dot{\Gamma}_p^-)^{-1}\|_\infty < \infty$ a.s. uniformly in $p = O((\log T)^a)$. In these evaluations it is essential that all bounds hold uniformly in the noise covariance contained in a compact set in the neighborhood of the true innovation variance since $\langle \varepsilon_t, \varepsilon_t \rangle$ will enter

the neighborhood a.s. for large T . This shows the claim of the lemma for $\hat{D} = D$. If D is estimated equation (3) shows that replacing \tilde{y}_t by \hat{y}_t does not change the uniform error bound on the estimated covariance sequence which is the only tool used in the proof given above. Hence the result also holds in this case.

If $p \geq -d \log T / (2 \log \rho_0)$ for some $d > 1$ it follows that $\rho^p = o(T^{-1/2})$ for $\rho > \rho_0$ sufficiently small proving the second claim of the lemma. \square

The next theorem is central to the consistency result and states an error bound on $\hat{\mathcal{K}}_p$ and \mathcal{K}_p using a specific normalization:

Lemma 3 *Let the assumptions of Theorem 1 hold. For a given multiindex j let a corresponding selector matrix $S_n \in \mathbb{R}^{ns \times n}$ be defined which is zero except for exactly one element in each column and at most one element in each row being equal to one where the position of the nonzero entries is indexed by j . Define the subset $M(n, j) \subset M(n)$ via the fact that $k(z) \in M(n, j)$ if $\mathcal{K}_n S_n \in \mathbb{R}^{n \times n}$ is nonsingular where \mathcal{K}_n corresponds to $k(z)$. Then there exists a finite set of multiindices J such that for each $j \in J$ the set $M(n, j)$ is open and dense in $M(n)$ and $\bigcup_{j \in J} M(n, j) = M(n)$. Further the restriction $\mathcal{K}_n S_n = I_n$ defines a canonical form on $M(n, j)$.*

For S_n corresponding to a multiindex j such that the true transfer function $k(z) \in M(n, j)$ let $S_p = [S'_n, 0]' \in \mathbb{R}^{ps \times n}$, $p \geq n$. Then for the estimate $\tilde{\mathcal{K}}_p = \hat{V}'_n (\hat{W}_p^-)^{-1}$ it holds that $\tilde{\mathcal{K}}_p S_p$ is nonsingular a.s. for T large enough and hence $\hat{\mathcal{K}}_p = [\tilde{\mathcal{K}}_p S_p]^{-1} \tilde{\mathcal{K}}_p$ is well defined a.s. Let $\mathcal{O}_f^\dagger = (\mathcal{O}'_f W \mathcal{O}_f)^{-1} \mathcal{O}'_f W$ where $W = (W_f^+)' W_f^+$ or $W = (\hat{\Gamma}_f^+)^{-1}$ in the CCA case. Then

$$\hat{\mathcal{K}}_p - \mathcal{K}_p = (\mathcal{O}'_f W \mathcal{O}_f)^{-1} \mathcal{O}'_f W (\hat{\beta} - \mathcal{O}_f \mathcal{K}_p) (I - S_p \mathcal{K}_p) + O(\max\{\rho^p, Q_T\}^2 f) \quad (6)$$

where \mathcal{O}_f and \mathcal{K}_p correspond to the representation of the true system implied by $\mathcal{K}_p S_p = I_n$. Thus for $p \geq -d \log T / (2 \log \rho_0)$ it follows that $\hat{\mathcal{K}}_p - \mathcal{K}_p = (\mathcal{O}'_f W \mathcal{O}_f)^{-1} \mathcal{O}'_f W (\hat{\beta} - \mathcal{O}_f \mathcal{K}_p) (I - S_p \mathcal{K}_p) + o(T^{-1/2}) = O(Q_T)$.

Proof: The first part of the theorem is straightforward to see noting that $\mathcal{K}_\infty = [K, (A - KC)K, \dots]$ is the reachability matrix corresponding to the inverse transfer function $k^{-1}(z) \in M(n)$ due to the stability and strict minimum-phase assumption. Then the theory developed for so called echelon overlapping forms (see Hannan and Deistler, 1988, Theorem 2.6.2) obviously has a dual counterpart operating on columns of the reachability matrix rather than on rows of \mathcal{O}_f . This follows since the transpose of a Hankel matrix is again a Hankel matrix and obviously the rank is not changed by taking the transpose. Then essentially Theorem 2.6.5, of Hannan and Deistler (1988) shows the claims. The fact that the restriction $\mathcal{K}_p S_p = I_n$ defines a unique representation on $M(n, j)$ is obvious.

Next $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p - \mathcal{O}_f \mathcal{K}_p = o(1)$ and the fact that $\mathcal{K}_p S_p$ is nonsingular by assumption show that there exists a random matrix $\hat{S}_T \in \mathbb{R}^{n \times n}$ such that $\hat{S}_T \hat{V}'_n \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{1/2} S_p$ is nonsingular a.s. for T large enough. A proof of the convergence in the case $f = p \rightarrow \infty$ at a certain rate can be found in Deistler *et al.* (1995) in the proof of Proposition 3.2. For the case of fixed f the arguments used there are straightforward to modify, also the different assumptions on the increase of p do not conflict with the proof as is easily verified. This shows that $\hat{\mathcal{K}}_p$ is welldefined for T large enough a.s. using this particular matrix S_p .

It remains to verify the order of convergence for $\hat{\mathcal{K}}_p - \mathcal{K}_p$ using the normalization $\hat{\mathcal{K}}_p S_p = \mathcal{K}_p S_p = I_n$. This will be done first for $\hat{D} = D$ and f is fixed and finite. Along the

lines of (Bauer *et al.*, 2000) a linearization of the SVD is derived in the following: Let $\hat{\mathcal{O}}_f^\dagger = (\hat{\mathcal{O}}_f'(\hat{W}_f^+)'\hat{W}_f^+\hat{\mathcal{O}}_f)^{-1}\hat{\mathcal{O}}_f'(\hat{W}_f^+)'\hat{W}_f^+$. Then

$$\begin{aligned}\hat{\mathcal{K}}_p - \mathcal{K}_p &= \hat{\mathcal{O}}_f^\dagger \hat{\beta} - \mathcal{O}_f^\dagger \mathcal{O}_f \mathcal{K}_p = \mathcal{O}_f^\dagger (\hat{\beta} - \mathcal{O}_f \mathcal{K}_p) + (\hat{\mathcal{O}}_f^\dagger - \mathcal{O}_f^\dagger) \hat{\beta} \\ &= \mathcal{O}_f^\dagger (\hat{\beta} - \mathcal{O}_f \mathcal{K}_p) P_{\mathcal{K}} + (\hat{\mathcal{O}}_f^\dagger - \mathcal{O}_f^\dagger) (\hat{\beta} - \mathcal{O}_f \mathcal{K}_p) P_{\mathcal{K}}\end{aligned}$$

where $P_{\mathcal{K}} = I_{ps} - S_p \mathcal{K}_p$ neglecting the dependence on p in the notation. In the third equality $(\hat{\mathcal{K}}_p - \mathcal{K}_p) S_p \mathcal{K}_p = 0$ and $\mathcal{K}_p P_{\mathcal{K}} = 0$ is used. Analogously

$$\hat{\mathcal{O}}_f - \mathcal{O}_f = \hat{\beta} \hat{\mathcal{K}}_p^\dagger - \mathcal{O}_f \mathcal{K}_p \mathcal{K}_p^\dagger = (\hat{\beta} - \mathcal{O}_f \mathcal{K}_p) \hat{\mathcal{K}}_p^\dagger + \mathcal{O}_f \mathcal{K}_p (\hat{\mathcal{K}}_p^\dagger - \mathcal{K}_p^\dagger)$$

where $\hat{\mathcal{K}}_p^\dagger = \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p' (\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p')^{-1}$ and \mathcal{K}_p^\dagger is defined analogously using \mathcal{K}_p and $\hat{\Gamma}_p^- = [\hat{\gamma}_{i-j}]_{i,j=1,\dots,p}$ where (i, j) -th entry is shown. From the uniform boundedness of the eigenvalues of $\hat{\Gamma}_p^-$ and $(\hat{\Gamma}_p^-)^{-1}$ and the analogous property of $\langle Y_{t,p}^-, Y_{t,p}^- \rangle$ and its inverse it follows using the structure of $\hat{\mathcal{K}}_p$ and \mathcal{K}_p that $(\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p')^{-1}$ and $(\mathcal{K}_p \hat{\Gamma}_p^- \mathcal{K}_p')^{-1}$ are of bounded norm a.s. $(\hat{\mathcal{O}}_f' \hat{W} \hat{\mathcal{O}}_f)^{-1}$ and $(\mathcal{O}_f' W \mathcal{O}_f)^{-1}$ are of bounded norm a.s. due to the assumptions on \hat{W}_f^+ and convergence of $\hat{\mathcal{O}}_f$ following from the convergence of $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p \rightarrow \mathcal{O}_f \mathcal{K}_p$, see above. Then the expression given above shows that $\hat{\mathcal{K}}_p - \mathcal{K}_p = O(p^{-1})$ since $\hat{\beta} - \mathcal{O}_f \mathcal{K}_p = O(p^{-1})$.

Next consider $\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p' - \mathcal{K}_p \hat{\Gamma}_p^- \mathcal{K}_p'$: Let k_T be such that $\hat{\mathcal{K}}_p - \mathcal{K}_p = O(k_T)$ where $p k_T \rightarrow 0$ which is possible from the discussion just provided. Further $\langle Y_{t,p}^-, Y_{t,p}^- \rangle - \hat{\Gamma}_p^- = O(Q_T)$ due to (2). Note that $\|\mathcal{K}_p\|_\infty < M$, $\|\hat{\Gamma}_p^-\|_\infty < M$, $p \in \mathbb{N}$, and $\|\mathcal{K}_p \hat{\Gamma}_p^-\|_\infty < M$, $p \in \mathbb{N}$ for some constant M a.s. since the typical entry of this matrix where $\hat{\Omega}$ is replaced by Ω (which does not change the arguments but makes the notation simpler) is $\mathbb{E}(x_t - (A - KC)^p x_{t-p}) y'_{t-j}$ for $1 \leq j \leq p$. Now $\mathbb{E} x_t y'_{t-j} = A^{j-1} \mathbb{E} x_{t+1} y'_t$ and $\mathbb{E} x_{t-p} y'_{t-j} = \mathbb{E} x_{t-p} x'_{t-p} (A^{p-j})'$ proving the claim.

Then according to the discussion below Lemma 1 it follows that $\hat{\mathcal{K}}_p^\dagger - \mathcal{K}_p^\dagger = O(\max(k_T, Q_T))$ and with f_T denoting the order of convergence of $\hat{\beta} - \mathcal{O}_f \mathcal{K}_p$ the expression for $\hat{\mathcal{O}}_f - \mathcal{O}_f$ given above implies that $\hat{\mathcal{O}}_f - \mathcal{O}_f = O(\max(f_T, k_T, Q_T))$. Here also $f_T p \rightarrow 0$ is used. Similarly the assumptions on \hat{W}_f^+ show that this implies that $\hat{\mathcal{O}}_f^\dagger - \mathcal{O}_f^\dagger = O(\max(f_T, k_T, Q_T))$. Inserting this into the expression given for $\hat{\mathcal{K}}_p - \mathcal{K}_p$ the boundedness of \mathcal{O}_f^\dagger implies $\hat{\mathcal{K}}_p - \mathcal{K}_p = O(\max(f_T, f_T k_T f, f_T Q_T f, f_T^2 f)) = O(k_T)$ by definition of k_T . This proves the second claim of the theorem for $\hat{D} = D$ and fixed finite f . Note that in the case $p \geq -d \log T / (2 \log \rho_0)$ we obtain $f_T = O(Q_T)$. Again only the uniform convergence of the estimated covariance sequence is used and hence the result also holds true for estimated \hat{D} according to equation (3). It remains to deal with the case $f = p \rightarrow \infty$. The proof given above in this situation holds unchanged up to the error bound on $\hat{\mathcal{O}}_f^\dagger - \mathcal{O}_f^\dagger$ where the bound on the error term has to be showed to hold uniformly in $f = O((\log T)^a)$. For the CCA weighting this is analogous to the arguments for $\hat{\mathcal{K}}_p^\dagger - \mathcal{K}_p^\dagger$ given above and hence the result also holds in this case. Note that in this case the inclusion of f in the order of convergence given above is necessary. \square

Note that the theorem states in all case that $\hat{\mathcal{K}}_p - \mathcal{K}_p = o(1)$. The next lemma states two facts that will frequently be used in the proof below:

Lemma 4 *Let the assumptions of Theorem 1 hold and let $\hat{\mathcal{K}}_p S_p = I_n$. Then*

$$\begin{aligned}\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p' - \mathcal{K}_p \hat{\Gamma}_p^- \mathcal{K}_p' &= O(\max(Q_T, \rho^p f)), \\ (A - KC)^p \langle x_{t,p}, z_t \rangle &= O(\rho^p)\end{aligned}$$

where z_t is any process such that $\langle z_t, z_t \rangle = O(1)$.

Proof: The first claim follows from repeated application of Lemma 1 as $\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p' - \mathcal{K}_p \dot{\Gamma}_p^- \mathcal{K}_p' = (\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle - \mathcal{K}_p \dot{\Gamma}_p^-) \hat{\mathcal{K}}_p' + \mathcal{K}_p \dot{\Gamma}_p^- (\hat{\mathcal{K}}_p - \mathcal{K}_p)'$. Note that the infinity norm of $\mathcal{K}_p \dot{\Gamma}_p^-$ is uniformly bounded (see the proof above). Then the discussion below Lemma 1 together with the uniform bound on the infinity norm $\dot{\Gamma}_p^-$ and $\dot{\Gamma}_p^- \mathcal{K}_p'$ and the fact that for $\hat{\mathcal{K}}_p - \mathcal{K}_p$ and $\langle Y_{t,p}^-, Y_{t,p}^- \rangle - \dot{\Gamma}_p^-$ the order of convergence multiplied by the number of entries tends to zero implies the first claim. The second claim is immediate from $(A - KC)^p = O(\rho^p)$ and $\langle x_{t-p}, z_t \rangle = O(1)$ which follows from a componentwise application of the Cauchy-Schwartz inequality noting that $\langle x_{t-p}, x_{t-p} \rangle = O(1)$ as is straightforward to show and $\langle z_t, z_t \rangle = O(1)$ by assumption. \square

In the following the representation (A, K, C) of the true system corresponds to the restriction $\mathcal{K}_p S_p = I_n$. The corresponding state will be denoted as x_t . The second step in the algorithm is the regression of \tilde{y}_t (assuming $\hat{D} = D$) onto \hat{x}_t in order to obtain an estimate

$$\begin{aligned} \hat{C} &= \langle \tilde{y}_t, \hat{x}_t \rangle \langle \hat{x}_t, \hat{x}_t \rangle^{-1} = \langle \varepsilon_t + C(A - KC)^p x_{t-p} + C \mathcal{K}_p Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p' (\hat{\mathcal{K}}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p')^{-1} \\ &= \langle \varepsilon_t + C \mathcal{K}_p Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p' (\mathcal{K}_p \dot{\Gamma}_p^- \mathcal{K}_p')^{-1} + O(\max(Q_T, \rho^p f)) \\ &= C \mathcal{K}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \hat{\mathcal{K}}_p' (\mathcal{K}_p \dot{\Gamma}_p^- \mathcal{K}_p')^{-1} + O(\max(Q_T, \rho^p f)) \\ &= C \mathcal{K}_p \langle Y_{t,p}^-, Y_{t,p}^- \rangle \mathcal{K}_p' (\mathcal{K}_p \dot{\Gamma}_p^- \mathcal{K}_p')^{-1} + O(\max(Q_T, \rho^p f)) = C + O(\max(Q_T, \rho^p f)). \end{aligned}$$

Here in the third equivalence the results of Lemma 4 is used, the fourth equivalence follows from $\langle \varepsilon_t, Y_{t,p}^- \rangle = O(Q_T)$ and the uniform bounds on $\|\hat{\mathcal{K}}_p\|_\infty$ and $\|(\mathcal{K}_p \dot{\Gamma}_p^- \mathcal{K}_p')^{-1}\|$ established above. The fifth follows from $\hat{\mathcal{K}}_p - \mathcal{K}_p = O(\max(Q_T, \rho^p f))$ according to Lemma 3. The final equality is due to $\langle Y_{t,p}^-, Y_{t,p}^- \rangle - \dot{\Gamma}_p^- = O(Q_T)$. Similar arguments show $\hat{A} - A = O(\max(Q_T, \rho^p f))$.

Next observe that

$$\begin{aligned} \hat{\Omega} &= \langle \tilde{y}_t - \hat{C} \hat{\mathcal{K}}_p Y_{t,p}^-, \tilde{y}_t - \hat{C} \hat{\mathcal{K}}_p Y_{t,p}^- \rangle \\ &= \langle \varepsilon_t + C(A - KC)^p x_{t-p} + (C \mathcal{K}_p - \hat{C} \hat{\mathcal{K}}_p) Y_{t,p}^-, \varepsilon_t + C(A - KC)^p x_{t-p} + (C \mathcal{K}_p - \hat{C} \hat{\mathcal{K}}_p) Y_{t,p}^- \rangle \\ &= \langle \varepsilon_t, \varepsilon_t \rangle + C(A - KC)^p \langle x_{t-p}, x_{t-p} \rangle [C(A - KC)^p]' \\ &\quad + (C \mathcal{K}_p - \hat{C} \hat{\mathcal{K}}_p) \langle Y_{t,p}^-, Y_{t,p}^- \rangle (C \mathcal{K}_p - \hat{C} \hat{\mathcal{K}}_p)' + O(\max(Q_T^2, \rho^{2p} f^2) p) \\ &= \hat{\Omega} + O(\max(Q_T^2, \rho^{2p} f^2) p) \end{aligned} \tag{7}$$

by using the above derived order bounds and Lemma 1 repeatedly. Further $\hat{x}_{t+1} = \hat{\mathcal{K}}_p Y_{t+1,p}^-$ and additionally

$$\langle Y_{t+1,p}^-, \hat{\varepsilon}_t \rangle = \langle Y_{t+1,p}^-, \tilde{y}_t - \hat{C} \hat{\mathcal{K}}_p Y_{t,p}^- \rangle = \langle Y_{t+1,p}^-, \varepsilon_t \rangle + \langle Y_{t+1,p}^-, Y_{t,p}^- \rangle (C \mathcal{K}_p - \hat{C} \hat{\mathcal{K}}_p)' + \langle Y_{t+1,p}^-, x_{t-p} \rangle (\bar{A}^p)' C'$$

where $\bar{A} = A - KC$. Each of these terms can be showed to be $O(\max(Q_T, \rho^p f))$ and it follows that replacing $\hat{\varepsilon}_t$ with ε_t introduces an error of magnitude $O(\max(Q_T, \rho^p f))$. This shows that

$$\begin{aligned} \hat{K} &= \langle \hat{x}_{t+1}, \hat{\varepsilon}_t \rangle \hat{\Omega}^{-1} = \langle \hat{x}_{t+1}, \varepsilon_t \rangle \hat{\Omega}^{-1} + O(\max(Q_T^2, \rho^{2p} f^2) p) \\ &= \hat{\mathcal{K}}_p \langle Y_{t+1,p}^-, \varepsilon_t \rangle \hat{\Omega}^{-1} + O(\max(Q_T^2, \rho^{2p} f^2) p) = \hat{\mathcal{K}}_p \langle Y_{t+1,p}^-, \varepsilon_t \rangle \hat{\Omega}^{-1} + O(\max(Q_T, \rho^p f)) \\ &= \hat{\mathcal{K}}_{p,1} \langle \tilde{y}_t, \varepsilon_t \rangle \hat{\Omega}^{-1} + O(\max(Q_T, \rho^p f)) = \hat{\mathcal{K}}_{p,1} \langle \varepsilon_t, \varepsilon_t \rangle \hat{\Omega}^{-1} + O(Q_T) = \hat{\mathcal{K}}_{p,1} + O(\max(Q_T, \rho^p f)) \end{aligned}$$

where $\hat{\mathcal{K}}_{p,1}$ denotes the first block column of $\hat{\mathcal{K}}_p$. In these evaluations we used the derivations given above in order to replace $\hat{\Omega}$ by $\dot{\Omega}$, $\hat{\varepsilon}_t$ by ε_t and the fact that $\langle \tilde{y}_{t-j}, \varepsilon_t \rangle = O(Q_T)$, $\langle x_t, \varepsilon_t \rangle = O(Q_T)$. This shows $\hat{K} - K = \hat{\mathcal{K}}_{p,1} - \mathcal{K}_{p,1} + O(\max(Q_T, \rho^p f)) = O(\max(Q_T, \rho^p f))$ using obvious notation. For $p \geq -d \log T / (2 \log \rho_0)$, $d > 1$ it follows that $\rho^p f = o(Q_T)$ for $\rho_0 < \rho$ small enough. Hence the error bound is of order $O(Q_T)$ in this case. The same bound for estimated \hat{D} again follows from (3).

It remains to show that there exists an index i such that the transformation to the overlapping echelon forms is differentiable and hence the rate of convergence holds also for the new coordinate system. Note that the multiindex i here is different from the index j used to define S_p in Lemma 2: There the index described columns of \mathcal{K}_p , now the index is used to describe rows of \mathcal{O}_f . The basic underlying idea is identical, however. To this end note that given any system representation the system representation in echelon coordinates is obtained from a state basis transformation using the transformation matrix $T = \tilde{S}_n \mathcal{O}_n$ where $\tilde{S}_n \in \mathbb{R}^{n \times ns}$ is a selector matrix selecting the rows of \mathcal{O}_n according to the multiindex i . Here i has to be selected such that T is nonsingular. Hence $(\hat{T} \hat{A} \hat{T}^{-1}, \hat{T} \hat{K}, \hat{C} \hat{T}^{-1})$ is equal to the system representation in the echelon overlapping form for $\hat{T} = \tilde{S}_n [\hat{C}', (\hat{C} \hat{A})', \dots, (\hat{C} \hat{A}^{n-1})']'$. Since $\hat{T} \rightarrow T$ according to the consistency for $(\hat{A}, \hat{K}, \hat{C})$ it follows that \hat{T} is nonsingular for large enough sample size proving the well definedness of $\hat{\theta}_i$. Furthermore consistency for $\hat{\theta}_i$ is immediate and the order of convergence follows from the order of convergence of $(\hat{A}, \hat{K}, \hat{C})$ and the obvious differentiability of \hat{T} as a function of the entries of $(\hat{A}, \hat{K}, \hat{C})$. This proves the theorem. \square

A.2 Proof of Theorem 2

Again the proof of the theorem consists in following the steps of the proof for the CLT in the case of conditionally homoskedastic innovations as is implicitly contained in Bauer and Ljung (2002). We will use linearization arguments analogously to the proof contained in (Bauer *et al.*, 1999). Consider the case of $\hat{D} = D$ first. Lemma 3 of (Bauer and Ljung, 2002) shows that ⁹

$$\text{vec}(\hat{A} - A, \hat{K} - K, \hat{C} - C) = \bar{M}_1 \text{vec}\langle \varepsilon_t, x_t \rangle + \bar{M}_{2,p} \text{vec}(\hat{\mathcal{K}}_p - \mathcal{K}_p) + o_P(T^{-1/2})$$

where both systems $(\hat{A}, \hat{K}, \hat{C})$ and (A, K, C) correspond to a particular coordinate neighborhood of the overlapping echelon forms. Part of the proof contained in (Bauer and Ljung, 2002) is also used in the proof of Theorem 1 given above. Since the main argument used in the proof is the uniform convergence of the estimated covariance sequence the proof of the lemma is unaltered (except for the exchange of all expressions involving γ_j by the identical expressions using $\dot{\gamma}_j$) under the assumptions on the noise of this paper with one exception: In Bauer and Ljung (2002) it is argued that $\sqrt{T} \langle \varepsilon_t - C(\hat{x}_t - x_t), \hat{x}_t \rangle$ converges in distribution, which has not been showed under the current assumptions. However, this is not needed since the proof that

$$\langle \varepsilon_t - C(\hat{x}_t - x_t), \hat{x}_t \rangle = (\hat{C} - C) \langle \hat{x}_t, \hat{x}_t \rangle + O(Q_T) = O(Q_T)$$

⁹To be precise the lemma shows the result for a different system representation. The change of the system representation, however, does not change the result as follows easily from using the Delta method. See the end of the proof of Theorem 1 for a related discussion.

is sufficient for the argument used in Bauer and Ljung (2002). Here the $O(Q_T)$ term in the middle expression is due to replacing \hat{y}_t by \tilde{y}_t . But $\hat{C} - C = O(Q_T)$ and $\langle \hat{x}_t, \hat{x}_t \rangle = O(1)$ has been showed in the proof of Theorem 1 given above. Therefore the conclusions of the lemma also hold in the setting of the current paper. For the construction of \bar{M}_1 and $\bar{M}_{2,p}$ see (Bauer and Ljung, 2002).

From equation (6) we obtain an expression for $\hat{\mathcal{K}}_p - \mathcal{K}_p$ whose central component is

$$\hat{\beta} - \mathcal{O}_f \mathcal{K}_p = \langle Y_{t,f}^+ - \mathcal{O}_f \mathcal{K}_p Y_{t,p}^-, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1}.$$

Further $Y_{t,f}^+ - \mathcal{O}_f \mathcal{K}_p Y_{t,p}^- = \mathcal{E}_f E_{t,f}^+ + \mathcal{O}_f (A - KC)^p x_{t-p}$. The second term does not contribute to the asymptotic distribution since $(A - KC)^p = o(T^{-1/2})$ and $\langle x_{t-p}, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} = O(1)$ under the assumptions of the theorem. According to lemma 1 we may replace $\langle Y_{t,p}^-, Y_{t,p}^- \rangle$ by $\bar{\Gamma}_p^-$ in the above expression without changing the asymptotic distribution since the difference is $o(T^{-1/2})$. Therefore the main ingredient in the asymptotic expression are terms of the form $\langle \varepsilon_{t+i}, \tilde{y}_{t-j} \rangle = \langle \varepsilon_t, \tilde{y}_{t-j-i} \rangle + o_P(T^{-1/2})$, $0 \leq i \leq f-1$, $1 \leq j \leq p$ using standard arguments (see also Lemma B.3 of Bauer, 2000). The same terms also appear in $\langle \varepsilon_t, x_t \rangle = \langle \varepsilon_t, Y_{t,p}^- \rangle \mathcal{K}'_p + o(T^{-1/2})$ using lemma 4.

The remainders of the proof are based on the arguments provided in Hannan and Deistler (1988), p. 145 ff, in particular on Bernstein's lemma (see e.g. Hannan and Deistler, 1988, Lemma 4.3.3.). In the notation of Hannan and Deistler (1988) we show that for the essential term derived above (using $r = f + p - 1$)

$$\text{vec}(\hat{A} - A, \hat{K} - K, \hat{C} - C) = M_r \text{vec}[\langle \varepsilon_t, \tilde{y}_{t-1} \rangle, \langle \varepsilon_t, \tilde{y}_{t-2} \rangle, \dots, \langle \varepsilon_t, \tilde{y}_{t-r} \rangle] + o_P(T^{-1/2})$$

we can use

$$\begin{aligned} x_T &:= \sqrt{T} M_r \text{vec}[\langle \varepsilon_t, \tilde{y}_{t-1} \rangle, \langle \varepsilon_t, \tilde{y}_{t-2} \rangle, \dots, \langle \varepsilon_t, \tilde{y}_{t-r} \rangle], \\ y_T(\varepsilon) &:= \sqrt{T} M_{r,1:m} \text{vec}[\langle \varepsilon_t, \tilde{y}_{t-1} \rangle, \langle \varepsilon_t, \tilde{y}_{t-2} \rangle, \dots, \langle \varepsilon_t, \tilde{y}_{t-m} \rangle], \\ z_T(\varepsilon) &:= \sqrt{T} M_{r,m+1:r} \text{vec}[\langle \varepsilon_t, \tilde{y}_{t-m-1} \rangle, \dots, \langle \varepsilon_t, \tilde{y}_{t-r} \rangle]. \end{aligned}$$

in order to deduce asymptotic normality of x_T and therefore of the system matrix estimates. Here m is a fixed integer, i.e. not depending on T . Further the notation $M_{r,a:b}$ is used to denote the matrix composed of the block columns indexed a up to b of M_r . Then for asymptotic normality of $\sqrt{T} \text{vec}(\hat{A} - A, \hat{K} - K, \hat{C} - C)$ it is sufficient to show that for each $\zeta > 0$, $\eta > 0$, $\varepsilon > 0$ there exists an m such that $y_T(\varepsilon)$ is asymptotically normal (where the asymptotic variance matrix $\Sigma(\varepsilon)$ of $y_T(\varepsilon)$ converges to Σ for $\varepsilon \rightarrow 0$) and $\mathbb{P}\{z_T(\varepsilon)' z_T(\varepsilon) > \zeta\} < \eta$. Using Bernstein's lemma again in order to rewrite \tilde{y}_{t-j} as a truncated sum of past ε_t 's plus truncation error shows that $\sqrt{T} \text{vec}[\langle \varepsilon_t, \tilde{y}_{t-1} \rangle, \langle \varepsilon_t, \tilde{y}_{t-2} \rangle, \dots, \langle \varepsilon_t, \tilde{y}_{t-m} \rangle]$ converges to a multivariate normal distribution under assumptions 1 on the noise. The arguments for this are provided by Lemma 4.3.4. of Hannan and Deistler (1988) and the discussion above the lemma. Therefore $M_{r,1:m} \rightarrow M_{1:m}$ for some matrix $M_{1:m}$ is a sufficient condition for $y_T(\varepsilon)$ to be asymptotically normal.

It remains to be showed that by choosing m sufficiently large $z_T(\varepsilon)$ can be made arbitrarily small (in probability). This will be done below. For a scalar sequence (i.e. $y_t \in \mathbb{R}$)

$$\mathbb{E} T \langle \varepsilon_t, \tilde{y}_{t-j} \rangle \langle \varepsilon_t, \tilde{y}_{t-j} \rangle = \frac{1}{T} \sum_{t,s=j+1}^T \mathbb{E} \varepsilon_t \varepsilon_s \tilde{y}_{t-j} \tilde{y}_{s-j} = \frac{1}{T} \sum_{t=j+1}^T \mathbb{E} \varepsilon_t^2 \tilde{y}_{t-j}^2 \leq (\mathbb{E} \varepsilon_t^4)^{1/2} (\mathbb{E} \tilde{y}_t^4)^{1/2} < \infty$$

due to the martingale difference assumption, strict stationarity and assumed finite fourth moments of ε_t . For vector valued ε_t the bound can be established for each coordinate separately. Note that the bound is uniform in $0 < j < T$. Hence

$$\begin{aligned} \mathbb{E}|\sqrt{T}x'M_{r,m+1:r}\text{vec}[\langle\varepsilon_t, \tilde{y}_{t-m-1}\rangle, \dots, \langle\varepsilon_t, \tilde{y}_{t-r}\rangle]| &\leq \sum_{j=m+1}^r \|M'_{r,j:j}\|_\infty \mathbb{E}\sqrt{T}\|\langle\varepsilon_t, \tilde{y}_{t-j}\rangle\|_1 \\ &\leq c \sum_{j=m+1}^r \|M'_{r,j:j}\|_\infty \left(\mathbb{E}\|\sqrt{T}\langle\varepsilon_t, \tilde{y}_{t-j}\rangle\|_2^2\right)^{1/2} \\ &\leq c\|M_{r,m+1:r}\|_\infty \end{aligned}$$

for some suitably chosen constant c (not necessarily the same in each inequality) and any vector x such that $\|x\|_\infty = 1$. Therefore in order to show that $\mathbb{P}\{z_T(\varepsilon)'z_T(\varepsilon) > \zeta\}$ can be made arbitrarily small it is sufficient to show that by choosing m large $\sup_{r \geq m} \|M_{r,m+1:r}\|_\infty$ can be made arbitrarily small. Then choosing m large enough

$$\mathbb{P}\{z_T(\varepsilon)'z_T(\varepsilon) > \zeta\} \leq \mathbb{P}\{\|z_T(\varepsilon)\|_\infty > c\sqrt{\zeta}\} \leq \mathbb{E}\|z_T(\varepsilon)\|_\infty / (c\sqrt{\zeta}) \leq \|M_{r,m+1:r}\|_\infty c' / \sqrt{\zeta} \leq \eta$$

for suitably small c using the equivalence of norms in finite dimensional spaces and the Markov inequality. Here c' denotes a suitable constant and the last inequality holds by choosing m large enough. Therefore it follows that a sufficient condition for both assumptions of Bernstein's lemma to hold is $\|[M_r, 0] - M_\infty\|_\infty \rightarrow 0$ for some matrix M_∞ with $\|M_\infty\|_\infty < \infty$. For the term $\sqrt{T}\langle\varepsilon_t, x_t\rangle = \sqrt{T}\langle\varepsilon_t, Y_{t,p}^-\rangle\mathcal{K}'_p + o_P(1)$ the condition is obvious. (Bauer and Ljung, 2002) establish the required condition for $\bar{M}_{2,p}$ in Lemma 3. There it is seen that $\bar{M}_{2,p}$ consists of a finite sum of terms of the form $N_j(L_{p,j} \otimes I_n)$ where N_j are finite dimensional matrices independent of the sample size and where $L_{p,j}$ is equal to either $[I_s, 0], \mathcal{K}_p\dot{\Gamma}_p^-$ or $\mathcal{K}_p[\mathcal{H}'_{1,p}, \dot{\Gamma}_p^-]$. For

$$\bar{M}_{2,p}\text{vec}(\hat{\mathcal{K}}_p - \mathcal{K}_p) = \bar{M}_{2,p} \left((I - \mathcal{K}'_p S'_p)(\dot{\Gamma}_p^-)^{-1} \otimes (\mathcal{O}'_f W \mathcal{O}_f)^{-1} \mathcal{O}'_f W \mathcal{E}_f \right) \text{vec} \left[\langle E_{t,f}^+, Y_{t,p}^- \rangle \right] + o_P(T^{-1/2})$$

the result follows from tedious but straightforward computations using the bound on the infinity norm of $(\dot{\Gamma}_p^-)^{-1}$ presented in Lemma 6.6.11 of (Hannan and Deistler, 1988) and the decomposition of $\bar{M}_{2,p}$ given above. Also the fact that $S'_p(\dot{\Gamma}_p^-)^{-1}$ converges to a matrix whose elements decrease exponentially is used which can be showed using the arguments in the proof of Lemma 5.5. of Bauer (2000). Here in the case that $f = p \rightarrow \infty$ Lemma 4.1.2. of (Bauer, 1998) is used in order to relate the expression using the matrix $\langle E_{t,f}^+, Y_{t,p}^- \rangle$ which apart from initial effects is block Hankel to an expression using only the essential entries, i.e. $\langle\varepsilon_t, \tilde{y}_{t-j}\rangle, j = 1, \dots, f + p - 1$. Also to account for $f \rightarrow \infty$ the fact that $\mathcal{O}'_f(\dot{\Gamma}_f^+)^{-1}$ converges to a matrix whose elements decrease exponentially is used (see equations (8,9,19) and Lemma 5.1. of Bauer (2000)).

The proof for the case of estimated \hat{D} is totally analogous noting that according to (3) the asymptotic distribution of the estimated covariance sequence does not change due to estimating D rather than knowing the true matrix. This concludes the proof. \square

A.3 Proof of Theorem 3

Note that in the statement of this theorem T_t is restricted to be a harmonic process. This has been done in order to preserve the equivalence of the estimators of θ as well as the parameters contained in D . For the pseudo maximum likelihood estimates it can be proved

using standard linearization techniques that the estimators $\hat{\theta}_{ML}$ and \hat{D}_{ML} are asymptotically equivalent to the estimators $\tilde{\theta}_{ML}$ obtained using the unfeasible process $\tilde{y}_t, t = 1, \dots, T$ as data for the estimation of θ and estimating D as $\hat{D} = \langle y_t, T_t \rangle \langle T_t, T_t \rangle^{-1}$. Here the meaning of asymptotically equivalent is that the difference of the estimators is of order $o_P(T^{-1/2})$. It will be shown below that $\tilde{\theta}_{ML} - \tilde{\theta}_{CCA} = o_P(T^{-1/2})$ where $\tilde{\theta}_{CCA}$ denotes the (unfeasible) CCA estimator based on $\tilde{y}_t, t = 1, \dots, T$. In the subspace case $\hat{D}_{CCA} = \hat{D}$ by definition. The theorem then follows from $\hat{\theta}_i - \tilde{\theta}_{CCA} = o_P(T^{-1/2})$ which is straightforward to verify using (3) reconsidering the proof of Theorem 2. Hence the details are omitted. Here all parameter vectors refer to the same multiindex i .

Therefore it remains to establish the asymptotic equivalence of the unfeasible estimators based on the process $(\tilde{y}_t)_{t \in \mathbb{Z}}$. Again the proof consists in paralleling the proof for the conditionally homoskedastic situation with analyzing the arguments which involve the assumption $\mathbb{E}\{\varepsilon_t \varepsilon_t' \mid \mathcal{F}_{t-1}\} = \Omega$ which has been used in Bauer (2000) but is not imposed in the present paper. The basis in this respect is the proof given in Bauer (2000). In the following we will step through the proof by noting the points where the arguments have to be changed rather than presenting a selfcontained proof. Such a proof would essentially replicate the arguments given in Bauer (2000) as will be clear from the following analysis. Instead we will only discuss the changes in the proof in order to hold also under the assumptions of this paper. This implies that the following proof can only be understood once Bauer (2000) has been studied. We also refrain from introducing concepts which are used in the proof but nowhere else in this paper. For details see Bauer (2000). All cited lemmas and theorems refer to Bauer (2000). The first three lemmas in appendix B are not specific to subspace methods. Lemma B.1 obviously continues to hold. In Lemma B.2. γ_j has to be replaced by $\hat{\gamma}_j$ in order for the inequalities to hold. Note that $\hat{\gamma}_j = O(1)$ continues to hold uniformly in $|j| < H_T$ due to ergodicity of ε_t , since $\langle \varepsilon_t, \varepsilon_t \rangle = \dot{\Omega} \rightarrow \Omega$ a.s. Lemma B.3 uses ergodicity, strict stationarity and equation (5.3.7) of Hannan and Deistler (1988) and hence also continues to hold. Lemma B.4 uses only $\langle \tilde{y}_{t-j}, \varepsilon_t \rangle = O(Q_T)$ which holds under the current assumptions and hence the proof holds unchanged.

Appendix A of Bauer (2000) describes the properties of the CCA estimate under the assumptions on f and p given in this paper. Lemma A.1 is implied by Theorem 1 as has been noted above. The proof of Lemma A.2. holds unchanged, since in no place $\hat{\gamma}_j - \gamma_j$ is used, but only $\hat{\gamma}_j = O(1)$. Lemma A.3. holds for the calculated backward system using Ω_0 , the true innovation variance, rather than an estimate thereof. If the backward system is calculated at $\dot{\Omega}$ the error bound on the backward system calculated on the basis of the estimated forward system cannot be showed with the tools provided above. Fortunately, this is not needed in the proof of the main result of Bauer (2000), see below for details. Lemma A.4 again holds unchanged, since it uses analogous arguments to Lemma A.2. Therefore all preliminary lemmas remain to hold under the less restrictive assumptions 1 (with the exception of the replacement of γ_j by $\hat{\gamma}_j$ for the uniform convergence argument).

In the main text of Bauer (2000) it is straightforward to verify that the arguments hold unchanged up to Lemma 5.1 given the fact that all preliminary lemmae hold with the appropriate changes indicated above. In Lemma 5.1 a matrix Z_f appears, which uses the true covariance sequence in its definition. We can introduce a matrix \dot{Z}_f converging to \dot{Z}_∞ such that the result given in the lemma still holds (a.s.): Here \dot{Z}_f is defined from the equation

$$\dot{\Xi}_f^{-1} \mathcal{O}_f = (\dot{\Gamma}_f^+)^{-1} \mathcal{O}_f + (\dot{\Gamma}_f^+)^{-1} \mathcal{O}_f (\dot{\Sigma}_x^{-1} - \mathcal{O}'_f (\dot{\Gamma}_f^+)^{-1} \mathcal{O}_f)^{-1} \mathcal{O}'_f (\dot{\Gamma}_f^+)^{-1} \mathcal{O}_f = (\dot{\Gamma}_f^+)^{-1} \mathcal{O}_f \dot{Z}_f$$

where all quantities correspond to the estimated system $(\hat{A}, \hat{K}, \hat{C})$ and $\hat{\Omega} = \langle \varepsilon_t, \varepsilon_t \rangle$ as the innovation variance. This follows since in the derivation of the results for Z_f the only assumption on the true covariance Ω used is the positive definiteness which also holds for $\hat{\Omega}$ a.s. for T large enough due to consistency. Note that $\hat{\Omega}(\hat{\theta}_{CCA})$ in Bauer (2000) is equal to $\langle \hat{\varepsilon}_t, \hat{\varepsilon}_t \rangle = \langle \varepsilon_t, \varepsilon_t \rangle + o(T^{-1/2})$ as follows from (7). This result is used in the proof of Lemma 5.2, where $Z_\infty \mathcal{K}_f^b(\hat{\theta}_{CCA})$ now has to be replaced by $\dot{Z}_\infty \dot{\mathcal{K}}_f^b(\hat{\theta}_{CCA})$ using $\dot{\mathcal{K}}_\infty^b(\hat{\theta}_{CCA}) = \mathcal{O}'_\infty(\dot{\Gamma}_\infty^+)^{-1}$. For definitions see Bauer (2000). The evaluations to show that the lemma holds with these replacements are identical to the ones given in Bauer (2000) and hence omitted.

The paragraph below Lemma 5.2 discusses the replacement of $\dot{\mathcal{K}}_f^b(\hat{\theta}_{CCA})$ by $\hat{\mathcal{O}}_f'(\hat{\Gamma}_f^+)^{-1}$. The discussion remains true since at no point the true covariance or conditional homoskedasticity is used.

Lemma 5.3 does not use conditional homoskedasticity and hence its conclusions remain true under the current set of assumptions. In Lemma 5.4. the error bound $\langle \hat{x}_t, \hat{y}_{t-p-1} \rangle = o(T^{-\varepsilon})$ is derived using the uniform convergence of sample covariances in combination with the fact $\mathbb{E}x_t \hat{y}'_{t-p-1} = o(T^{-\varepsilon})$ for suitable $\varepsilon > 0$. This bound is derived from $x_t = A^p x_{t-p} + \sum_{j=0}^{p-1} A^j K \varepsilon_{t-j}$ where $A^p = o(T^{-\varepsilon})$ for p obeying the lower bound. Therefore replacing the true covariances with the dotted quantities leaves the error bound unchanged. Note that

$$\dot{\mathcal{K}}_f^b(\hat{\theta}_{CCA}) = \hat{\mathcal{O}}_f'(\hat{\Gamma}_f^+)^{-1} + O(Q_T) = \mathcal{K}_f^b(\hat{\theta}_{CCA}^b) + O(Q_T)$$

where the first equality has been stated above. The second equality is the backward analogue to $\hat{\mathcal{K}}_p - \mathcal{K}_p(\hat{\theta}_{CCA}) = O(Q_T)$ according to Lemma 5.1 and the proof of Lemma A.1. Here $\hat{\theta}_{CCA}^b$ denotes a parameter vector corresponding to the estimated backward system (see Bauer, 2000, for a definition). Then the proof of Lemma 5.4. follows from

$$(\dot{\mathcal{K}}_f^b(\hat{\theta}_{CCA}) - \mathcal{K}_f^b(\hat{\theta}_{CCA}^b))(\hat{\mathcal{O}}_f - \mathcal{O}_f(\hat{\theta}_{CCA})) = O(Q_T^2 p) = o(T^{-1/2})$$

using the error bound on $\dot{\mathcal{K}}_f^b(\hat{\theta}_{CCA}) - \mathcal{K}_f^b(\hat{\theta}_{CCA}^b)$ derived above, $\hat{\mathcal{O}}_f - \mathcal{O}_f(\hat{\theta}_{CCA}) = O(Q_T)$ by Lemma 5.1. and the expression of $\hat{\mathcal{O}}_f - \mathcal{O}_f$ given in the proof of Theorem 1 above in order to apply Lemma 1. This finishes the proof for derivatives with respect to entries in K .

Corresponding to the derivation concerning derivatives with respect to entries in A the same reasoning holds as the previously established lemmatae are used with the indicated changes (i.e. exchanging Ω for $\hat{\Omega}$ in the formulations where needed). For derivatives with respect to entries in C the proof of Lemma 5.5 follows the same lines as before, where in each occurring covariance matrix $\hat{\Omega}$ is used in place of Ω . This concludes the proof of the Theorem. \square

A.4 Proof of Theorem 4

The main argument in the proof of Theorem 3 in Bauer (2001) is $\|\hat{X}_{f,p} - X_0\|_2 = O(Q_T \sqrt{fp})$, where $\hat{X}_{f,p} = \hat{W}_f^+ \hat{\beta} \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{1/2}$ and X_0 denotes the same quantity corresponding to the true covariances. By replacing true covariances with the corresponding dotted quantities, i.e. using $\dot{X}_0 = W_f^+ \mathcal{O}_f \mathcal{K}_p(\dot{\Gamma}_p^-)^{1/2}$ the same result holds in the present framework from noting that $\hat{W}_f^+ - W_f^+ = O(Q_T)$ according to assumptions 2, $\hat{\beta} - \mathcal{O}_f \mathcal{K}_p = O(Q_T)$ as shown above and $\langle Y_{t,p}^-, Y_{t,p}^- \rangle - \dot{\Gamma}_p^- = O(Q_T)$ due to the uniform convergence of the covariance estimates. Here also the fact that the difference of the square root of two positive definite matrices is of the same order of magnitude as the difference in the matrices themselves (for small deviations) as is showed e.g. in Bauer (1998). Note that the rank of \dot{X}_0 does not depend on the noise

covariance matrix $\hat{\Omega}$ but only on the rationality of $k(z)$. This is the only change in the proof of the theorem as compared to Theorem 3 of (Bauer, 2001). Therefore we refer to the original article for details. \square .

A.5 Proof of Theorem 5

Deal first with the case $\hat{D} = D$ where the effects of T_t are assumed to be known. The estimation of D will be dealt with later. In order to simplify notation let $E_t = \text{vech}[\varepsilon_t \varepsilon_t']$ and $\hat{E}_t = \text{vech}[\hat{\varepsilon}_t \hat{\varepsilon}_t']$ where $\hat{\varepsilon}_t$ here denotes the estimates of the innovations sequence based on the Kalman filter corresponding to the subspace estimate $\hat{\theta}$ using zero initial conditions. Further let $E_{t,p} = \text{vech}[\varepsilon_{t-1} \varepsilon_{t-1}', \dots, \varepsilon_{t-p} \varepsilon_{t-p}']$ and again analogously $\hat{E}_{t,p}$ is defined. Here 'vech' denotes the vector of the stacked lower triangular part of a matrix or the vector of the stacked lower triangular parts of a set of matrices. Let $x_t = [1, E_{t,p}]'$, $\hat{x}_t = [1, \hat{E}_{t,p}]'$ and $\beta \in \mathbb{R}^{[s(s+1)/2] \times [ps(s+1)/2+1]}$. Then the two estimation equations can be written as

$$E_t = \beta x_t + u_t, \quad \hat{E}_t = \beta \hat{x}_t + \hat{u}_t$$

The IV estimates of β based on the two data sets hence are given as

$$\begin{aligned} \hat{\beta}_{IV} &= \langle E_t, z_t \rangle \hat{W} \langle z_t, x_t \rangle \left(\langle x_t, z_t \rangle \hat{W} \langle z_t, x_t \rangle \right)^{-1} \\ \tilde{\beta}_{IV} &= \langle \hat{E}_t, z_t \rangle \hat{W} \langle z_t, \hat{x}_t \rangle \left(\langle \hat{x}_t, z_t \rangle \hat{W} \langle z_t, \hat{x}_t \rangle \right)^{-1}. \end{aligned}$$

A typical choice for \hat{W} would be $\hat{W} = \langle z_t, z_t \rangle^{-1}$ which in a conditionally homoskedastic framework corresponds to the optimal choice. In the theorem this weighting is assumed to be chosen. The proof below will show that more general weightings could be used. Note that $\langle z_t, z_t \rangle \rightarrow \mathbb{E} z_t z_t' > 0$, $\langle x_t, z_t \rangle \rightarrow \mathbb{E} x_t z_t'$, $\langle E_t, z_t \rangle \rightarrow \mathbb{E} E_t z_t'$ due to ergodicity, stationarity and finite moments. Furthermore

$$\sqrt{T} \text{vec}(\hat{\beta}_{IV} - \beta) = (\Sigma_{XZ} \otimes I) \frac{1}{\sqrt{T}} \sum_{t=1}^T z_t \otimes u_t + o_P(1)$$

for $\hat{W} = \langle z_t, z_t \rangle$ where $\Sigma_{XZ} = (\mathbb{E} x_t z_t' (\mathbb{E} z_t z_t')^{-1} \mathbb{E} z_t x_t')^{-1} \mathbb{E} x_t z_t' (\mathbb{E} z_t z_t')^{-1}$. Note that for $u_t = E_t - \beta x_t$ the process $z_t \otimes u_t$ is a strictly stationary ergodic square integrable martingale difference such that $T^{-1} \sum_{t=1}^T z_t z_t' \otimes u_t u_t' \rightarrow \mathbb{E} z_t z_t' \otimes u_t u_t'$ (again due to ergodicity and existence of the expectation in the limit). Therefore $\sqrt{T}^{-1} \sum_{t=1}^T z_t \otimes u_t$ converges in distribution to a Gaussian limit (see e.g. Davidson, 1994, Theorem 24.3). Furthermore the variance of the limiting Gaussian variable of $\sqrt{T} \text{vec}(\hat{\beta}_{IV} - \beta)$ is equal to

$$[\Sigma_{XZ} \otimes I] \mathbb{E} (z_t z_t' \otimes u_t u_t') [\Sigma_{XZ}' \otimes I]$$

It is straightforward to verify that sufficient conditions for the equivalence of the asymptotic properties of $\hat{\beta}_{IV}$ and $\tilde{\beta}_{IV}$ are (\xrightarrow{P} denoting convergence in probability)

$$\sqrt{T} \langle E_t - \hat{E}_t, z_t \rangle \xrightarrow{P} 0, \quad \sqrt{T} \langle x_t - \hat{x}_t, z_t \rangle \xrightarrow{P} 0.$$

Both $x_t - \hat{x}_t$ and $E_t - \hat{E}_t$ involve expressions composed of entries of lagged values of $\varepsilon_t \varepsilon_t' - \hat{\varepsilon}_t \hat{\varepsilon}_t'$. Note that

$$\hat{\varepsilon}_t - \varepsilon_t = \hat{\varepsilon}_t(\hat{\theta}) - \hat{\varepsilon}_t(\theta_0) + o(\rho^t) = \partial \hat{\varepsilon}_t(\bar{\theta})(\hat{\theta} - \theta_0) + o(\rho^t)$$

using a mean value expansion (∂ denoting partial derivative with respect to the entries in θ) where the $o(\rho^t)$ term corresponds to neglecting the initial state and $0 < \rho_0 < \rho < 1$. As usual $\bar{\theta}$ denotes a value on the line segment between θ_0 and $\hat{\theta}$. Here $\hat{\varepsilon}_t(\theta)$ denotes the innovations estimated based on the Kalman filter corresponding to the parameter vector θ assuming initial state $x_1 = 0$. Since $\hat{\varepsilon}_t(\theta) = y_t - DT_t - C(\theta)x_t(\theta)$ it follows that $\partial\hat{\varepsilon}_t(\theta)$ is \mathcal{F}_{t-1} measurable and hence $\mathbb{E}\{\varepsilon_t\partial_i\hat{\varepsilon}_t(\theta)' \mid \mathcal{F}_{t-1}\} = 0$ for all i , where ∂_i denotes partial derivative with respect to the i -th component of θ . This conditional uncorrelatedness is the reason for the equivalence of the asymptotic distributions: Since z_t is assumed to be \mathcal{F}_{t-p-1} measurable the statement above implies that $\mathbb{E}\{\varepsilon_{t-j}\partial_i\hat{\varepsilon}_{t-j}(\theta)'z_{t,r} \mid \mathcal{F}_{t-p-1}\} = 0, j = 0, 1, \dots, p$ for each coordinate $z_{t,r}$ of z_t and hence the instruments are uncorrelated with the highest order term in the estimation error $\hat{E}_t - E_t$ and $\hat{E}_{t,p} - E_{t,p}$. Note that $\sqrt{T}(\hat{\theta} - \theta_0)$ converges in distribution according to Theorem 2.

In the following we will only deal with the scalar case. The multivariate case is only notationally more complex. The main arguments are identical. Hence consider (using $\varepsilon_t = o(t^{1/2}), \partial\hat{\varepsilon}_t(\bar{\theta}) = o(t^{1/2})$ due to convergence of sample second moments)

$$\begin{aligned}\hat{\varepsilon}_t^2 - \varepsilon_t^2 &= (\hat{\varepsilon}_t - \varepsilon_t)(\hat{\varepsilon}_t + \varepsilon_t) = 2\varepsilon_t(\hat{\varepsilon}_t - \varepsilon_t) + (\hat{\varepsilon}_t - \varepsilon_t)^2 \\ &= 2\varepsilon_t\partial\hat{\varepsilon}_t(\bar{\theta})(\hat{\theta} - \theta_0) + \partial\hat{\varepsilon}_t(\bar{\theta})(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'\partial\hat{\varepsilon}_t(\bar{\theta})' + o(\rho^t t^{1/2})\end{aligned}$$

implying

$$\begin{aligned}\sqrt{T}\langle\hat{E}_t - E_t, z_t\rangle &= \frac{1}{\sqrt{T}}\sum_{t=1}^T\left(2\varepsilon_t\partial\hat{\varepsilon}_t(\bar{\theta})(\hat{\theta} - \theta_0) + \partial\hat{\varepsilon}_t(\bar{\theta})(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'\partial\hat{\varepsilon}_t(\bar{\theta})'\right)z_t' + o(1) \\ &= \frac{1}{T}\sum_{t=1}^T 2\varepsilon_t z_t' \partial\hat{\varepsilon}_t(\bar{\theta})\left(\sqrt{T}(\hat{\theta} - \theta_0)\right) + \sqrt{T}(\hat{\theta} - \theta_0)'\partial\hat{\varepsilon}_t(\bar{\theta})'\partial\hat{\varepsilon}_t(\bar{\theta})(\hat{\theta} - \theta_0)z_t' + o(1)\end{aligned}$$

where the $o(1)$ term is due to neglecting initial effects which follows from standard arguments. Hence $\langle\varepsilon_t\partial\hat{\varepsilon}_t(\bar{\theta})', z_t\rangle = o_P(1)$ and $\langle\partial\hat{\varepsilon}_t(\bar{\theta})', \partial\hat{\varepsilon}_t(\bar{\theta})'z_{t,i}\rangle = O_P(1)$ (here $z_{t,i}$ denotes the i -th coordinate of z_t) are sufficient conditions for $\sqrt{T}\langle E_t - \hat{E}_t, z_t\rangle \rightarrow 0$ in probability.

Consider the term $\langle\varepsilon_t\partial\hat{\varepsilon}_t(\bar{\theta})', z_t\rangle = T^{-1}\sum_{t=1}^T\varepsilon_t\partial\hat{\varepsilon}_t(\bar{\theta})'z_t'$ first. Note that $\varepsilon_t\partial_i\hat{\varepsilon}_t(\bar{\theta})'z_t'$ is uncorrelated to $\varepsilon_s\partial_j\hat{\varepsilon}_s(\bar{\theta})'z_s'$ for $s \neq t$ and each pair i, j which can be seen conditioning on $\mathcal{F}_{\max(s,t)-1}$. Further the variance of each term is bounded uniformly in $\bar{\theta}$ in a compact neighborhood of θ_0 which can be seen as follows: Note that $\partial\hat{\varepsilon}_t(\bar{\theta})' = \Theta_t(\bar{\theta})Y_{t,t-1}^-$ where $\Theta_t(\bar{\theta}) = [\Theta_{t,i}(\bar{\theta})]_{i=1,\dots,t-1}, \Theta_{t,i}(\bar{\theta}) \in \mathbb{R}^{2ns \times s}$. For θ in a compact neighborhood of θ_0 we obtain $\|\Theta_{t,i}(\bar{\theta})\| \leq C\rho^i$ for some $\rho_0 < \rho < 1$. This follows from the fact that the steady state Kalman filter has the state space representation $(A(\theta) - K(\theta)C(\theta), K(\theta), -C(\theta))$. It is then easy to see that also the derivative has a state space representation with stable A -matrix. Then consistency for $\hat{\theta}$ shows that $\bar{\theta}$ enters the compact neighborhood a.s. A uniform bound on the variance of $\partial\varepsilon_t(\theta)$ is then easily obtained from $\|\Gamma_\infty^-\|_\infty < \infty$ (see Lemma 2). It is then simple to show that $\mathbb{E}\langle\varepsilon_t\partial_i\hat{\varepsilon}_t(\bar{\theta})', z_t\rangle'(\langle\varepsilon_t\partial_i\hat{\varepsilon}_t(\bar{\theta})', z_t\rangle) \rightarrow 0$ for each i implying in probability convergence of $\langle\varepsilon_t\partial\hat{\varepsilon}_t(\bar{\theta})', z_t\rangle$.

Next consider $\langle\partial\hat{\varepsilon}_t(\bar{\theta})', \partial\hat{\varepsilon}_t(\bar{\theta})'z_{t,i}\rangle$. Note that due to the norm bound on z_t it follows that $|z_{t,i}| < C$ for some constant $C < \infty$. Hence

$$-\langle\partial\hat{\varepsilon}_t(\bar{\theta})', \partial\hat{\varepsilon}_t(\bar{\theta})'\rangle C \leq \langle\partial\hat{\varepsilon}_t(\bar{\theta})', \partial\hat{\varepsilon}_t(\bar{\theta})'z_{t,i}\rangle \leq \langle\partial\hat{\varepsilon}_t(\bar{\theta})', \partial\hat{\varepsilon}_t(\bar{\theta})'\rangle C$$

in the sense that the difference is a positive definite matrix. The convergence and boundedness of this term follows from Theorem 2.1. of Findley *et al.* (2001). Here the uniform bounds on

the norm of $\Theta_{t,j}$ and their exponential decrease derived above provide the main argument. The result $\sqrt{T}\langle x_t - \hat{x}_t, z_t \rangle \rightarrow 0$ in probability can be showed analogously noting that the entries of $x_t - \hat{x}_t$ are in the scalar case (except for the first component) equal to $\hat{\varepsilon}_{t-i}^2 - \varepsilon_{t-i}^2, i = 1, \dots, p$. Combining these two results leads to the proof of the theorem for the case $\hat{D} = D$. In the case that D is estimated prior to applying CCA the arguments change slightly. In that case the estimation error $\hat{\varepsilon}_t - \varepsilon_t$ contains the additional term $(\hat{D} - D)T_t$. Since $(T_t)_{t \in \mathbb{Z}}$ is assumed to be independent of $(\varepsilon_t)_{t \in \mathbb{Z}}$ it follows from ergodicity that $\langle \varepsilon_t T_t, z_t \rangle \rightarrow 0$ and $\langle T_t, T_t z_{t,i} \rangle = O(1)$ follows as above from the boundedness of z_t .

It remains to examine the estimation of the asymptotic variance matrix: The unfeasible estimator based on the knowledge of ε_t equals

$$\left[\hat{\Sigma}_{XZ} \otimes I \right] \left[\frac{1}{T} \sum_{t=p+1}^T (z_t z_t' \otimes \hat{u}_t \hat{u}_t') \right] \left[\hat{\Sigma}'_{XZ} \otimes I \right]$$

where $\hat{\Sigma}_{XZ} = (\langle x_t, z_t \rangle \langle z_t, z_t \rangle^{-1} \langle z_t, x_t \rangle)^{-1} \langle x_t, z_t \rangle \langle z_t, z_t \rangle^{-1}$ and $\hat{u}_t = E_t - \hat{\beta}_{IV} x_t$ denotes the residuals. Now $\langle z_t, z_t \rangle \rightarrow \mathbb{E} z_t z_t'$ due to ergodicity and the existence of the moments. Similarly $\langle z_t, x_t \rangle \rightarrow \mathbb{E} z_t x_t'$ follows. Hence the assumptions on the rank of $\mathbb{E} x_t z_t'$ implies that $\hat{\Sigma}_{XZ} \rightarrow \Sigma_{XZ}$. Ergodicity together with the finiteness of the expectation ensures that $T^{-1} \sum_{t=p+1}^T z_t z_t' \otimes x_t x_t' \rightarrow \mathbb{E} z_t z_t' \otimes x_t x_t'$ and $T^{-1} \sum_{t=p+1}^T z_t z_t' \otimes E_t E_t' \rightarrow \mathbb{E} z_t z_t' \otimes E_t E_t'$. Since $u_t = E_t - \beta x_t$ is a linear combination of E_t and x_t also $T^{-1} \sum_{t=p+1}^T z_t z_t' \otimes u_t u_t' \rightarrow \mathbb{E} z_t z_t' \otimes u_t u_t'$ follows. Note that $u_t u_t'$ contains fourth powers of ε_t . Replacing u_t by its estimates $\hat{u}_t = E_t - \hat{\beta}_{IV} x_t$ does not change the limit due to consistency of $\hat{\beta}_{IV}$ as is straightforward to show.

The corresponding feasible estimator using \hat{E}_t and \hat{x}_t is of the form

$$\left[\tilde{\Sigma}_{XZ} \otimes I \right] \left[\frac{1}{T} \sum_{t=p+1}^T (z_t z_t' \otimes \tilde{u}_t \tilde{u}_t') \right] \left[\tilde{\Sigma}'_{XZ} \otimes I \right]$$

where $\tilde{\Sigma}_{XZ} = (\langle \hat{x}_t, z_t \rangle \langle z_t, z_t \rangle^{-1} \langle z_t, \hat{x}_t \rangle)^{-1} \langle \hat{x}_t, z_t \rangle \langle z_t, z_t \rangle^{-1}$ and $\tilde{u}_t = \hat{E}_t - \tilde{\beta}_{IV} \hat{x}_t$ denotes the residuals. Since it has been showed above that $\sqrt{T}\langle \hat{x}_t - x_t, z_t \rangle \rightarrow 0$ in probability it follows that $\langle \hat{x}_t, z_t \rangle \rightarrow \mathbb{E} x_t z_t'$ in probability and therefore $\tilde{\Sigma}_{XZ} \rightarrow \Sigma_{XZ}$. The arguments to show that $T^{-1} \sum_{t=p+1}^T z_t z_t' \otimes \tilde{u}_t \tilde{u}_t' \rightarrow \mathbb{E} z_t z_t' \otimes u_t u_t'$ are identical to the proof given above: Note that

$$\tilde{u}_t - u_t = \hat{E}_t - \tilde{\beta}_{IV} \hat{x}_t - E_t + \beta x_t = \hat{E}_t - E_t - (\tilde{\beta}_{IV} - \beta_{IV}) \hat{x}_t - \beta (\hat{x}_t - x_t).$$

Hence using the mean value expansion for $\hat{E}_t - E_t$ discussed above e.g. $T^{-1} \sum_{t=p+1}^T z_t z_t' \otimes u_t (\hat{E}_t - E_t)' \rightarrow 0$ can be showed analogously to $\sqrt{T}\langle \hat{x}_t - x_t, z_t \rangle \xrightarrow{p} 0$, the evaluations now being simpler since normalization is by T^{-1} rather than only $1/\sqrt{T}$. The same arguments also apply for the remaining terms. This concludes the proof of the Theorem. \square