Consistency of System Identification by Global Total Least Squares

submitted to the Journal of Time Series Analysis

C. Heij [*]   W. Scherrer [†]

June 5, 1996

Abstract

Global total least squares (GTLS) is a method for the identification of linear systems where no distinction between input and output variables is required. This method has been developed within the deterministic behavioural approach to systems. In this paper we analyse statistical properties of this method when the observations are generated by a multivariable stationary stochastic process. In particular, sufficient conditions for the consistency of GTLS are derived. This means that, when the number of observations tends to infinity, the identified deterministic system converges to the system that provides an optimal appoximation of the data generating process. The two main results are the following. GTLS is consistent if a guaranteed stability bound can be given a priori. If this information is not available, then consistency is obtained (at some loss of finite sample efficiency) if GTLS is applied to the observed data extended with zero values in past and future.

Keywords: Linear systems, stochastic systems, behavioural approach, factor analysis, estimation, consistency.

[*]Econometric Institute and Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, phone +31-10-4081277, fax +31-10-4527746, e-mail heij@ect.few.eur.nl.

[†]Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Argentinierstraße 8, A-1040 Vienna, Austria, phone +43-1-58801-4444, fax +43-1-5054524, e-mail wo_sch@e119ws1.tuwien.ac.at.

1   Introduction

System identification is concerned with the determination of sufficiently sim-
ple models that give a sufficiently accurate description of the observed data.
Identification methods differ in the specification of the model class and in the
way the complexity and accuracy of models is evaluated. For our purposes it
is helpful to distinguish models according to their treatment of the observed
variables, that is, models can be (i) closed or open, and (ii) symmetric or
asymmetric. We call a model closed if it specifies the behaviour of all the
observed variables, and open if it leaves some variables unexplained. A model
is symmetric if the *a priori* assumptions are the same for all the observed
variables, and asymmetric if this is not the case. Most of the approaches
that have been developed for the estimation of multivariable systems con-
sider either closed symmetric models or open asymmetric ones. To be more
specific, within the time series literature one usually models the variables as
a jointly stationary process, which corresponds to a closed symmetric model.
For example, ARMA models describe the observed process w in terms of a
white noise process ε by means of square polynomial matrices $P(z^{-1}), Q(z^{-1})$
as

$$P(z^{-1})w = Q(z^{-1})\varepsilon \tag{1}$$

where z^{-1} denotes the shift operator defined by $(z^{-1}w)(t) := w(t-1)$. From
the standard textbooks on identification along these lines we mention Hannan
(1970), Anderson(1971), Priestley (1981) and Lütkepohl (1993). On the
other hand, within the control engineering literature the usual starting point
consists of input-output relations. These models are open, as the behaviour of
the inputs is left unspecified, and asymmetric, as they require prior selection
of input and output variables. A well-known example is prediction error
identification, of which ARMAX modelling forms a special case. If the inputs
are denoted by u, the outputs by y and the noise by ε, then these models are
described in terms of polynomial matrices P, Q, N, with P and Q square, as

$$P(z^{-1})y = N(z^{-1})u + Q(z^{-1})\varepsilon \tag{2}$$

The asymmetry of this model is especially clear in the modelling of the
disturbances, as one usually assumes that these act on the outputs alone
and not on the inputs, that is, u and ε are assumed to be uncorrelated. For

this and related approaches we refer to Davis and Vinter (1985), Ljung(1987), Caines (1988) and Hannan and Deistler (1988).

In practice it may not always be clear which variables act as inputs and which ones as outputs, or what are the properties of the disturbances. This is particularly relevant if the disturbances are seen as a result of model inaccuracies rather than as the effect of physical noise acting on the system. Such situations demand a more symmetric treatment of the system variables. On the other hand, it may not be realistic to ask for a closed system model if the available information is not sufficient to formulate relations for all the system variables. In this case one needs methods for the identification of open symmetric models. Several approaches have been developed for this purpose, in particular errors-in-variables models, see Deistler (1989), Beghelli et al. (1990) and Deistler and Scherrer (1992), and system behaviours, see Willems (1986, 1991), Heij (1989) and Roorda (1995a). The first approach has the advantage that it treats stochastic systems which allows a statistical analysis of identification methods, whereas behaviours have been developed within a deterministic setting. A disadvantage of errors-in-variables models is that identification requires rather strong noise assumptions, whereas identification within the behavioural framework is a matter of deterministic approximation. In order to make a statistical analysis of identification procedures possible without the need for strong prior assumptions on the noise, a synthesis of the two approaches is proposed in Heij et al. (1995) in terms of dynamic factor models. Hereby it is assumed that the observations are generated by a stationary stochastic process, denoted by w. A factor model is a decomposition

$$w = \hat{w} + \tilde{w} \tag{3}$$

$$R(z^{-1})\hat{w} = 0 \tag{4}$$

where $\hat{w}$ and $\tilde{w}$ are stationary processes and $R(z^{-1})$ is a polynomial matrix with less rows than columns. Here $\hat{w}$ is called the latent process, and $\tilde{w}$ is the corresponding error process. As errors are allowed in all components this gives a symmetric treatment of the variables, and as the number of restrictions on $\hat{w}$ is less than the number of observed variables this model is also open. Factor models are evaluated in terms of their complexity, measured by the number of degrees of freedom of the latent process $\hat{w}$, and in terms of the magnitude of the error process $\tilde{w}$. For the identification of these models

from observed data we consider in this paper the so-called method of global total least squares (GTLS), see Roorda (1995a) and Roorda and Heij (1995).

The statistical properties of identification methods for ARMA and AR-MAX models are well-established. In particular, sufficient conditions for consistency and asymptotic normality of estimators have been derived, see Hannan (1970), Anderson(1971), Ljung(1987), Caines (1988) and Hannan and Deistler (1988). In this paper we make a first step in the statistical analysis of the estimation of open symmetric models, by considering the consistency of global total least squares for the identification of factor models. Stated more precisely, assume that the tolerated complexity of factor models has been fixed. A model is called optimal if it has minimal total least squares error under this complexity constraint, that is, if it minimizes the error $\{E\|\tilde{w}(t)\|^2\}^{1/2}$ where $\|\cdot\|$ is the Euclidean norm. So this corresponds to the best achievable approximation $\hat{w}$ of the process w in case this process was fully known. In practice, the available information consists of an observed time series. An identification method is called consistent if the identified system converges to the optimal system in case the number of observations tends to infinity. The two main results of this paper are, roughly stated, the following. The global total least squares method is consistent if a guaranteed stability bound for the optimal model can be given *a priori*. If this information is not available, then consistency is obtained (at some loss of finite sample efficiency) if GTLS is applied to the observed data extended with zero values in past and future. This paper extends earlier partial results that were presented in Heij et al. (1995) and Heij and Scherrer (1994, 1995).

The paper has the following structure. In Section 2 we specify the data generating process, the class of factor models and the identification method. This is done in terms of system behaviours, that is, free of parametrization. For algorithmic purposes, the representation of systems by means of so-called isometric state space models is described in Section 3. The results on consistency are presented in Section 4, and they are illustrated by two simple simulation examples in Section 5. Section 6 concludes, and proofs are in the Appendix.

2 Linear Systems and Factor Models

In this section we describe the main elements of the identification problem considered in this paper, that is, the data generating process, the model class, and the identification procedure. In order to streamline the exposition we will make several simplifying assumptions, some of which could be relaxed without affecting the results. For a more general treatment of some of the issues discussed in this section we refer to Willems (1986, 1991), Heij et al. (1995) and Roorda (1995a).

2.1 The Data Generating Process

Let the observed process be denoted by w, and the number of observed variables by q. A symmetric treatment of the variables requires that the prior assumptions on the data generating process are also of a symmetric nature. We will assume that the data are generated by a stationary stochastic process in discrete time $\mathsf{Z} := \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. To be more precise, we assume that w is a purely nondeterministic, weakly stationary process of full rank, with zero mean and finite second order moments. The full rank condition means that the process satisfies no (linear) deterministic restrictions, and the condition of being purely nondeterministic means in practice that possible purely harmonic components have been removed. These assumptions are usual in the analysis of stationary time series. Under the above assumptions the process has a standardized Wold representation

$$w = T(z^{-1})\varepsilon \qquad (5)$$

where ε is a q-dimensional white noise process with zero mean and unit covariance matrix. Here $T(z^{-1}) = \sum_{k=0}^{\infty} T_k z^{-k}$ is a causal transfer function with causal inverse, with T_0 invertible and $\sum_{k=0}^{\infty} \|T_k\|^2 < \infty$. This also means that the process has a spectral density given by $\Sigma = \frac{1}{2\pi} TT^*$, where T^* is the adjoint defined by $T^*(z) := T'(z^{-1})$ with T' the transposed of T. Throughout this paper we will impose the condition that the transfer function T is absolutely summable, $\sum_{k=0}^{\infty} \|T_k\| < \infty$, so that the spectrum is a bounded function on the unit circle. Such processes are called linear, and they are ergodic in the sense that the sample covariances almost surely converge pointwise to the process covariances if the number of observations

tends to infinity, see Hannan (1970, Theorem IV.6). In addition we assume that $\mathrm{E}\{\varepsilon(t) \mid \varepsilon(s), s \leq t - 1\} = 0$, that $\mathrm{E}\{\varepsilon(t)\varepsilon'(t) \mid \varepsilon(s), s \leq t - 1\} = \mathrm{E}\,\varepsilon(t)\varepsilon'(t) = I$, and that the fourth order moments of the process ε exist. All the above assumptions are satisfied for Gaussian ARMA processes, but they hold also true under weaker conditions.

2.2   The Model Class

The purpose of identification is to estimate the main characteristics of the data generating process from observed data. In order to define our model class, we first give a behavioural description of deterministic linear systems. This approach was introduced by Willems (1986) and gives a symmetric treatment of the system variables, in contrast with the usual description in terms of input-output systems. The behaviour of a deterministic system with q variables is defined as the set of all trajectories $\hat{w} : \mathsf{Z} \to \mathsf{R}^q$ that may arise within the restrictions imposed by the system. So a behaviour is a subset $\mathcal{B}$ of $(\mathsf{R}^q)^{\mathsf{Z}}$. We consider behaviours that are linear, time invariant, and complete. This means that $\mathcal{B} \subset (\mathsf{R}^q)^{\mathsf{Z}}$ is a linear subspace that is invariant under the shift operator z^{-1}, defined by $(z^{-1}\hat{w})(t) := \hat{w}(t - 1)$, and that the behaviour is in addition closed in the topology of pointwise convergence. The last condition means that for a sequence $\hat{w}_n \in \mathcal{B}$ which converges pointwise (in R^q) to $\hat{w}_0 \in (\mathsf{R}^q)^{\mathsf{Z}}$ there holds that also $\hat{w}_0 \in \mathcal{B}$. This may seem a somewhat technical condition, but it means that the behaviour corresponds to a linear, time invariant, finite dimensional system. The observed variables w can be partitioned into inputs and outputs so that the behaviour consists of all input-output trajectories that can be generated by a linear, time invariant, finite dimensional system with freely chosen initial conditions. In the sequel we will simply use the term linear system to refer to a linear, time invariant, complete behaviour.

Definition 1 *A* linear system *is a linear, shift invariant subset $\mathcal{B} \subset (\mathsf{R}^q)^{\mathsf{Z}}$ that is closed in the topology of pointwise convergence.*

A linear system is called controllable if every past trajectory in $\mathcal{B}$ can be driven in finite time into every future trajectory in $\mathcal{B}$.

Every linear system can be represented in polynomial form, as the solution set of the polynomial equations $R(z^{-1})\hat{w} = 0$ as in equation (4). The

representation of a given system by a polynomial matrix is highly non-unique, but every representation has the same (polynomial) rank, say $p = \text{rank}(R)$. Then (4) defines a finite dimensional input-output system with p outputs and $m := q - p$ inputs. We denote by n the number of states (initial conditions) of the system. This number equals the minimally achievable sum of the lags of the p equations in (4).

Our model class consists of factor models, which are defined in terms of linear systems as follows. A factor model of an observed process w is a decomposition (3) where the latent process $\hat{w}$ satisfies the polynomial equations (4). More precisely, almost all realizations of $\hat{w}$ should satisfy the equations (4). For simplicity we will assume throughout that the processes $w, \hat{w}$ and $\tilde{w}$ are stationary and purely nondeterministic.

For comparison, the traditional model of static factor analysis is of the form $w = Lf + \tilde{w}$, where w is a vector of observed variables, f a lower dimensional vector of unobserved factor variables, $\tilde{w}$ a vector of unobserved noise components, and L a matrix of factor loadings. If we define the latent variables by $\hat{w} = Lf$, then these variables satisfy deterministic linear equations as the matrix L does not have full row rank. The model (3), (4) is the dynamic version of this model, where the latent process satisfies deterministic linear difference equations. The interpretation of this model is that the data generating process w approximately satisfies the restrictions of the linear system $\mathcal{B}$, at the expense of an error $\tilde{w}$. Imposing stronger restrictions on the latent process $\hat{w}$ will in general result in a larger error $\tilde{w}$. So factor models involve a trade-off between complexity and goodness of fit.

The error of a factor model (3), (4) is defined in terms of the variance of the noise process by $\{\text{E} \|\tilde{w}(t)\|^2\}^{1/2}$. In practical applications it may be relevant to stress the importance of certain variables or certain frequency regions, which can easily be achieved by appropriate prefiltering of the data. The error of a system $\mathcal{B}$ for the process w is defined as the smallest achievable error of all factor models (3) where the latent process satisfies the equations (4) of the system $\mathcal{B}$. The optimal factor model is given by $\hat{w} = P_{\mathcal{B}}w$ and $\tilde{w} = (I - P_{\mathcal{B}})w$, where $P_{\mathcal{B}}$ is the operator of orthogonal projection (for square summable trajectories) onto the linear system $\mathcal{B}$, see Heij et al. (1995). In terms of the spectrum Σ of the data generating process, the error of this

factor model is given by

$$e(\mathcal{B}) = \left[\int_{-\pi}^{\pi} \text{trace}\{(I - P_{\mathcal{B}}(e^{i\omega}))\Sigma(e^{i\omega})\}d\omega \right]^{1/2}. \qquad (6)$$

The complexity of a system $\mathcal{B}$ is defined as the pair (m, n), with m the number of inputs and n the number of states of the system. The complexity measures the number of degrees of freedom that are present in the latent process $\hat{w}$, as the dimension of the system $\mathcal{B}$ restricted to a time interval of length $N \geq n$ is given by $Nm + n$.

Our aim is to find systems with minimal error for a given a priori bound on the complexity. Such systems are called optimal. We are mainly interested in Pareto optimal systems, that is, among the systems of optimal fit we prefer the ones with minimal complexity. We use a partial ordering of complexities where (m, n) is less complex than (m', n') if both $m \leq m'$ and $n \leq n'$. Every linear system can be decomposed as $\mathcal{B} = \mathcal{B}_c + \mathcal{B}_a$ where $\mathcal{B}_c$ is the largest controllable system contained in $\mathcal{B}$ and where $\mathcal{B}_a$ is a finite dimensional set. The system $\mathcal{B}_c$ is called the controllable part of $\mathcal{B}$, and $\mathcal{B}_a$ is an autonomous system without inputs so that the trajectories in $\mathcal{B}_a$ are completely determined by the initial conditions. As $\mathcal{B}_a$ contains no non-zero square summable trajectories it follows that $P_{\mathcal{B}} = P_{\mathcal{B}_c}$ so that $e(\mathcal{B}) = e(\mathcal{B}_c)$. If $\mathcal{B}$ is not controllable then $\mathcal{B}_c$ has complexity (m, n') with $n' < n$. It follows that Pareto optimal systems are controllable. The problem of determining optimal models for given spectrum Σ and complexity (m, n) will be discussed in Section 3.

We summarize the foregoing in a definition.

Definition 2   *(i) A factor model of an observed process w is a decomposition $w = \hat{w} + \tilde{w}$, with $\hat{w}$ and $\tilde{w}$ stationary and where $\hat{w}$ satisfies equations (4).*

  (ii) The complexity of a system $\mathcal{B}$ is given by the pair (m, n), with m the number of inputs and n the number of states.

 (iii) The error of a system $\mathcal{B}$ with respect to a process w is defined by (6), where Σ is the spectrum of w and $P_{\mathcal{B}}$ the orthogonal projection onto $\mathcal{B}$.

 (iv) For given complexity (m, n), a system is optimal if it minimizes the error (6) under this complexity constraint, and Pareto optimal if in addition all less complex systems have strictly larger error.

2.3 The Identification Method

We now consider the situation where the process w is unknown, and the available information consists of an observed time series generated by the process. That is, the data $w_N = w(\omega)_{|[1,N]}$ consists of a realization $w(\omega)$ of the process observed on a time interval of length N. For the identification of a linear system on the basis of these data we minimize the global total least squares (GTLS) distance, defined by

$$e_N(\mathcal{B}) = \min\{\frac{1}{\sqrt{N}}\|w_N - w_N^a\| \text{ with } w_N^a \in \mathcal{B}_{|[1,N]}\} \tag{7}$$

where $\|\cdot\|$ is the Euclidean norm in (qN)-dimensional space. This distance involves the total squares, in the sense that approximations in all the variables are allowed. It is also global in the sense that the approximation w_N^a should not only locally satisfy the system equations (4), as in prediction oriented criteria, but also globally because the full trajectory w_N^a should satisfy the laws of the behaviour $\mathcal{B}$.

For given complexity (m, n), a system is optimal for the observed data if it minimizes the error (7) under this complexity constraint. We mention that, in contrast with the process error (6), Pareto optimal systems for the GTLS criterion (7) need not be controllable.

In applications it is of importance to make a balanced trade-off between complexity and goodness of fit. This can be achieved by varying the complexity and evaluating the corresponding variations in the goodness of fit of optimal models. In this paper we do not further consider the issue of complexity specification, and we take the complexity as given. For fixed complexity (m, n) we denote the optimal model for the data w_N by $\mathcal{B}_N^*$. This is a random system, as it depends on the data that are generated by the stochastic process w. Let $\mathcal{B}^*$ be the optimal system in the sense of minimizing (6) under the complexity constraint, that is, the system with minimally achievable error if the data generating process is known. The quality of identification methods can be measured in terms of the discrepancy between the identified system and the optimally achievable result, that is, the distance between $\mathcal{B}_N^*$ and $\mathcal{B}^*$. The basic question considered in this paper is whether global total least squares is a consistent identification method. That is, we investigate in which sense and under which conditions it holds true that

$$\lim_{N \to \infty} \mathcal{B}_N^* = \mathcal{B}^*. \tag{8}$$

In order to analyse this question we need to consider the algorithms for the computation of $\mathcal{B}^*$ and $\mathcal{B}_N^*$ in more detail, and this is the topic of the next section.

3   Parametrization and Optimization

In the foregoing section we defined systems and their complexity and goodness of fit in behavioural terms, that is, in terms of the set of trajectories compatible with the system restrictions. For computational purposes we need a parametric representation of linear systems. One possibility is to use polynomial representations (4). However, for the GTLS error (7) it has proved convenient to use a more structured type of representation, by means of so-called isometric state space models. In this section we describe this parametrization and its use in GTLS identification, and for further background we refer to Roorda(1995a, 1995b) and Roorda and Heij (1995).

3.1   Isometric State Models

Every linear system can be represented in terms of state variables $\hat{x}$ and driving variables $\hat{v}$ by means of the equations

$$\hat{x}(t+1) = A\hat{x}(t) + B\hat{v}(t) \tag{9}$$

$$\hat{w}(t) = C\hat{x}(t) + D\hat{v}(t). \tag{10}$$

In contrast with the usual input-state-output model, here all observed variables are seen as outputs of a system driven by auxiliary forces. So the model is symmetric, and it is also open because the driving forces are unrestricted so that not all components of $\hat{w}$ are explained. This can also be interpreted as a dynamic factor model, where the observed variables $\hat{w}$ are generated by the factors $\hat{x}$ that evolve over time and with additional unrestricted factors $\hat{v}$. For a given system this representation is highly non-unique. A system of complexity (m, n) can be represented by m driving variables and n state variables, and not by a smaller number of these factor variables. Such representations are called minimal. If (A, B, C, D) is a minimal representation

of a system then all its minimal representations are given by the feedback group $(S(A + BF)S^{-1}, SBR, (C + DF)S^{-1}, DR)$ where S and R are $n \times n$ and $m \times m$ invertible matrices and F is an arbitrary $m \times n$ matrix. In terms of parameters, the minimality of representations amounts to the conditions that the $n \times (n + m)$ matrix $[A \quad B]$ has full row rank n, the $q \times m$ matrix D has full column rank m, and the matrix pair $(A + BF, C + DF)$ is observable for all $m \times n$ matrices F.

In our approach we will not incorporate all linear systems in the model class, as we will require that the systems are stabilizable. In behavioural terms, a system is called stabilizable if all trajectories on finite time intervals admit a continuation within the system that converges to zero. In parametric terms, a system is stabilizable if there exists a matrix F such that all eigenvalues of the matrix $A + BF$ are contained in the open unit disc. So we exclude systems that are not stabilizable. Pareto optimal systems for the process error (6) are controllable, and such systems are also stabilizable.

Stabilizable systems can be represented by means of isometric state space models, where the $(n + q) \times (n + m)$ matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{11}$$

is isometric, that is, it has orthogonal columns of unit length. Minimal isometric representations are unique up to block-unitary transformations. That is, if a behaviour has minimal isometric representation (A, B, C, D) then all such representations are given by (UAU', UBV, CU', DV) with U and V $n \times n$ and $m \times m$ unitary matrices. For fixed dimensions (m, n) we define the parameter set $\Theta \subset \mathsf{R}^{n \times n} \times \mathsf{R}^{n \times m} \times \mathsf{R}^{q \times n} \times \mathsf{R}^{q \times m}$ by

$$\Theta = \{(A, B, C, D); \begin{pmatrix} A & B \\ C & D \end{pmatrix}' \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I_n & O \\ O & I_m \end{pmatrix}\} \tag{12}$$

So the parameter set consists of isometric representations, but note that minimality is not required. By $\mathcal{B}(\theta)$ we denote the linear system corresponding to $\theta \in \Theta$, that is, all the trajectories w that can be generated by the equations (9), (10) for these values of the parameters. This parametrization of systems is not injective, due to the non-uniqueness of isometric representations. The image of this parametrization is the set of all stabilizable systems with complexity (m, n') with $n' \leq n$.

3.2   Computation of Model Errors

The behavioural error of factor models is defined by $e(\theta) := e(\mathcal{B}(\theta))$ given in (6), and the GTLS error by $e_N(\theta) := e_N(\mathcal{B}(\theta))$ as in (7). For later purposes we need more explicit expressions, as these are the objective functions that are minimized in identification. For given $\theta = (A, B, C, D)$ let $\tilde{B}$ and $\tilde{D}$ be $n \times (q - m)$ and $q \times (q - m)$ matrices such that the matrix

$$\begin{pmatrix} A & B & \tilde{B} \\ C & D & \tilde{D} \end{pmatrix} \tag{13}$$

is unitary, i.e., square and isometric. Now $\theta = (A, B, C, D)$ is a minimal representation if and only if $[A \ \ B]$ has full row rank and $(A, \tilde{B})$ is controllable. The system $\mathcal{B}(\theta)$ is controllable if and only if (A, B) is controllable. We define the transfer functions $\hat{G}(z^{-1}) = D + \sum_{k=1}^{\infty} CA^{k-1}Bz^{-k}$ and $\tilde{G}(z^{-1}) = \tilde{D} + \sum_{k=1}^{\infty} CA^{k-1}\tilde{B}z^{-k}$, and the adjoints by $\hat{G}^*(z^{-1}) := \hat{G}'(z)$ and $\tilde{G}^*(z^{-1}) := \tilde{G}'(z)$. It follows from the isometry condition (12) that $A'A + C'C = I_n$, so that A is a stable matrix and the eigenspaces corresponding to eigenvalues on the unit circle are not observable. This means that $\hat{G}$ and $\tilde{G}$ are bounded rational transfer functions, so that $\hat{v} := \hat{G}^*w$ and $\tilde{v} := \tilde{G}^*w$ are well-defined stationary processes, as well as $\hat{w} = \hat{G}\hat{v}$ and $\tilde{w} = \tilde{G}\tilde{v}$. There holds that $\hat{G}^*\hat{G} = I_m$ and $\tilde{G}^*\tilde{G} = I_{q-m}$, so that the transfer functions are both isometric, and they are orthogonal as $\hat{G}^*\tilde{G} = 0$. Further, $P = \hat{G}\hat{G}^*$ is the operator of orthogonal projection onto the system $\mathcal{B}(\theta)$ and $I_q - P = \tilde{G}\tilde{G}^*$ onto its orthogonal complement. This gives the following parametric expression for the error (6), where the subindex θ denotes the dependence on the parameters.

$$e(\theta) = \{E\|\tilde{v}_\theta(t)\|^2\}^{1/2} = \left[\int_{-\pi}^{\pi} \mathrm{trace}\{\tilde{G}_\theta^*(e^{i\omega})\Sigma(e^{i\omega})\tilde{G}_\theta(e^{i\omega})\}d\omega\right]^{1/2} \tag{14}$$

Although for given parameters θ the extension with $(\tilde{B}, \tilde{D})$ in (13) is only determined up to right multiplication by a unitary matrix, the expression (14) only depends on θ and not on the chosen extension. For the actual computation of the above error it is convenient to use the following recursive algorithm in terms of state space models. For a given process w and parameters $\theta = (A, B, C, D)$ with $(\tilde{B}, \tilde{D})$ a unitary extension as in (13), the

optimal approximation $\hat{w} = \hat{G}\hat{G}^*w$ of w within the system $\mathcal{B}(\theta)$ and the corresponding error $\tilde{w} = \tilde{G}\tilde{G}^*w$ can be obtained as follows.

$$x(t) = A'x(t+1) + C'w(t) \tag{15}$$

$$\hat{v}(t) = B'x(t+1) + D'w(t) \tag{16}$$

$$\tilde{v}(t) = \tilde{B}'x(t+1) + \tilde{D}'w(t) \tag{17}$$

$$\hat{x}(t+1) = A\hat{x}(t) + B\hat{v}(t), \quad \hat{w}(t) = C\hat{x}(t) + D\hat{v}(t) \tag{18}$$

$$\tilde{x}(t+1) = A\tilde{x}(t) + \tilde{B}\tilde{v}(t), \quad \tilde{w}(t) = C\tilde{x}(t) + \tilde{D}\tilde{v}(t) \tag{19}$$

These recursions can also be used for the computation of the GTLS error (7), as follows. If data are observed over the time interval $[1, N]$, then compute (15), (16), (17) backwards for $t = N, N-1, \dots, 1$ and determine the end state $x(N+1)$ by minimizing $\frac{1}{N}\sum_{t=1}^{N} \|\tilde{v}(t)\|^2$. The resulting $\hat{v}$ and $\tilde{v}$ are the inputs to (18), (19) with initial conditions $\hat{x}(1) = x(1)$ obtained from (15) and $\tilde{x}(1) = 0$. If θ is a minimal representation, then the time series $\hat{w}$ generated by (18) is the optimal approximation $w_N^a \in \mathcal{B}(\theta)_{|[1,N]}$ in (7) with corresponding GTLS error $e_N(\theta) = \{\frac{1}{N}\sum_{t=1}^{N} \|\tilde{w}(t)\|^2\}^{1/2} = \{\frac{1}{N}\sum_{t=1}^{N} \|\tilde{v}(t)\|^2\}^{1/2}$. So the difference between the approximation of a given process in (6) and the approximation of finite observed data in (7) consists in the determination of the states $x(N+1), \hat{x}(1), \tilde{x}(1)$. However, if θ is not minimal then the computed time series $\hat{w}$ need not belong to $\mathcal{B}(\theta)_{|[1,N]}$, since the state $\hat{x}(1) = x(1)$ may be non-reachable within the system $\mathcal{B}(\theta)$. This means that in this case we only obtain a lower bound, as $e_N(\theta) \geq e_N^*(\theta)$ where $e_N^*(\theta) = \{\frac{1}{N}\sum_{t=1}^{N} \|\tilde{v}(t)\|^2\}^{1/2}$ with $\tilde{v}$ obtained as before by means of (15), (17) with optimally chosen end state $x(N+1)$ so as to minimize $\{\frac{1}{N}\sum_{t=1}^{N} \|\tilde{v}(t)\|^2\}^{1/2}$.

We summarize the main points of the foregoing discussion. References for the proofs are given in the Appendix.

Proposition 1 *(i) Every stabilizable linear system can be represented in isometric state space form (9), (10), (12), and in identification we take as parameter set Θ in (12) with fixed values for (m, n).*

(ii) The set of systems parametrized by Θ is the set of all stabilizable systems of complexity (m, n') with $n' \leq n$, and $n' = n$ if θ is minimal.

(iii) On the process level, the error (6) is given in parametric terms by $e(\theta) = \{E\|\tilde{v}(t)\|^2\}^{1/2}$ in (14) and this can be computed by (15), (17).

(iv) *The GTLS error (7) satisfies $e_N(\theta) \geq e_N^*(\theta)$ obtained from (15), (17) with optimally chosen end state $x(N + 1)$, and equality holds true if θ is minimal.*

3.3   Structural Properies of the Identification Problem

In order to determine optimal models of complexity at most (m, n), the objective functions e and e_N should be minimized over the parameter set Θ. This is a highly nonlinear problem, and a closed-form solution is not available. Gauss-Newton algorithms have been developed that converge to local minima of the identification criteria (6), (7), see Roorda (1995b). These algorithms directly apply to the GTLS criterion (7) for finite observed data. On the process level, the system error (6) can be expressed as the GTLS problem over infinite time for the standardized Wold transfer function T in (5), see Heij et al. (1995). Here we will not consider further algorithmical details, and instead we describe some more general properties of the optimization problems at hand. Apart from the parameter set Θ, we consider in the sequel also the subsets Θ_ρ, with $0 < \rho \leq 1$, defined by

$$\Theta_\rho = \{\theta \in \Theta; \lambda_{max}(A) \leq \rho\} \tag{20}$$

where $\lambda_{max}(A)$ denotes the maximum modulus of the eigenvalues of the matrix A. Because of the isometry condition (12) there holds that $\Theta_1 = \Theta$. Further, if $\lambda_{max}(A) = 1$ then this representation is not minimal, that is, the corresponding system $\mathcal{B}(\theta)$ has less than n active state variables. So, for a given system there always exists an isometric representation with $\lambda_{max}(A) < 1$. The restriction of the parameter set Θ_ρ is that this gives a guaranteed stability bound, in the sense that the effect of initial conditions in (15), (18), (19) dies out at an exponential rate of at least ρ.

The following result states some properties that are useful for the minimization of e and e_N over Θ or Θ_ρ. Here we consider the parameter sets as subsets of the Euclidean space $\mathsf{R}^{(n+q)(n+m)}$ with the usual topology.

Proposition 2   *(i) Θ and Θ_ρ are compact;*

 (ii) e is continuous on Θ;

 (iii) e_N is continuous at minimal points $\theta \in \Theta$ if $N \geq n$;

(iv) for fixed N and $\theta_k \to \theta$ there holds $\limsup_{k\to\infty} e_N(\theta_k) \le e_N(\theta)$.

The proof of this and other results can be found in the Appendix. This shows that the identification problem corresponds to the minimization of relatively well-behaved functions on compact domains. This is further investigated in the following section, where we also consider the question of consistency (8), that is, the question whether the (global) minima of e_N converge to those of e if the number of observations N tends to infinity.

4   Consistency

Stated in general terms, an identification method is called consistent if the model identified from finite data converges to an optimal approximation of the data generating process if the number of observations tends to infinity. We analyse this question for GTLS on two levels, that is, on the parametric level in terms of the state space representation (9), (10), and on the behavioural level of linear systems as formulated in (8). Throughout this section we assume that the maximal tolerated complexity (m, n) is fixed. Most of the analysis will be in parametric terms, as this is most close to the algorithmic implementation of the GTLS procedure, see (15) - (19). Further we will distinguish two cases, that is, modelling with a guaranteed stability bound given by the parameter set Θ_ρ in (20), and modelling over the full parameter set Θ when the observed data are extended with zeros in past and future. In the last case the criterion (7) is replaced by the GTLS approximation over infinite time of the time series $(\ldots, 0, 0, w_N, 0, 0, \ldots)$. In terms of the Fourier transform $w_N^f := \frac{1}{\sqrt{N}} \sum_{t=1}^N w(t) e^{-i\omega t}$ and the periodogram $S_N = \frac{1}{2\pi} w_N^f (w_N^f)^*$ this GTLS error is given by

$$e_N^0(\theta) = \left[\int_{-\pi}^{\pi} \operatorname{trace}\{\tilde{G}_\theta^*(e^{i\omega}) S_N(e^{i\omega}) \tilde{G}_\theta(e^{i\omega})\} d\omega \right]^{1/2}. \tag{21}$$

4.1   Identification with Guaranteed Stability Bound

We first state some auxiliary results and introduce some notation. Let $\hat{G}_\theta$ and $\tilde{G}_\theta$ denote the isometric transfer functions corresponding to θ defined in Section 2.3, and let $\tilde{v}_\theta = \tilde{G}_\theta^* w$ and $\tilde{w}_\theta = \tilde{G}_\theta \tilde{G}_\theta^* w$. By $\tilde{V}_N(\theta)$ and $\tilde{W}_N(\theta)$ we denote the vectors of values of $\tilde{v}_\theta(t)$ and $\tilde{w}_\theta(t)$ respectively, stacked

in reverse order $t = N, N - 1, \ldots, 1$. Further, by $\Pi_N(\theta)$ we denote the $N(q - m) \times N(q - m)$ matrix of orthogonal projection onto the image of the $N(q - m) \times n$ matrix $(\tilde{B}, A\tilde{B}, \ldots, A^{N-1}\tilde{B})'$. Finally, a class of (possibly non-causal) filters $\{F_\theta = \sum_{k=-\infty}^{\infty} F_\theta(k) z^{-k}; \theta \in \Theta\}$ is called uniformly stable if $\sum_{k=-\infty}^{\infty} \{\sup_\theta \|F_\theta(k)\|\} < \infty$.

Lemma 3   *(i) For every $\rho < 1$ the classes of filters $\{\tilde{G}_\theta^*; \theta \in \Theta_\rho\}$ and $\{P_\theta = \tilde{G}_\theta \tilde{G}_\theta^*; \theta \in \Theta_\rho\}$ are uniformly stable, but this does not hold true for $\rho = 1$.*

(ii) The GTLS error (7) satisfies

$$\frac{1}{N}\tilde{W}_N'(\theta)\tilde{W}_N(\theta) \geq \{e_N(\theta)\}^2 \geq \{e_N^*(\theta)\}^2 = \frac{1}{N}\tilde{V}_N'(\theta)\{I - \Pi_N(\theta)\}\tilde{V}_N(\theta)$$

$$(22)$$

Lemma 4 *For every $\rho < 1$ and for $N \to \infty$, the following convergence results hold true almost surely and uniformly over Θ_ρ :*

(i) $\frac{1}{N}\tilde{V}_N'(\theta)\tilde{V}_N(\theta) \to e^2(\theta)$ and $\frac{1}{N}\tilde{W}_N'(\theta)\tilde{W}_N(\theta) \to e^2(\theta)$;

(ii) $\frac{1}{N}\tilde{V}_N'(\theta)\Pi_N(\theta)\tilde{V}_N(\theta) \to 0$;

(iii) $e_N(\theta) \to e(\theta)$.

We now consider the consistency of GTLS over the parameter set Θ_ρ, with $\rho < 1$ fixed. For fixed complexity (m, n) we denote by $\Theta_\rho^* = \operatorname{argmin}_{\{\theta \in \Theta_\rho\}} e(\theta)$ the set of optimal parameters for the process and by $\Theta_{\rho,N}^* = \operatorname{argmin}_{\{\theta \in \Theta_\rho\}} e_N(\theta)$ the set of parameters of GTLS models for the observed data. We consider Θ in a natural way as subset of $\mathsf{R}^{(n+q)(n+m)}$ with Euclidean norm.

Theorem 5 *In parametric terms, GTLS is consistent over the parameter set Θ_ρ for every $\rho < 1$, as for $N \to \infty$ there holds that almost surely*

$$\sup_{\{\theta_N \in \Theta_{\rho,N}^*\}} \inf_{\{\theta \in \Theta_\rho^*\}} \|\theta_N - \theta\| \to 0.$$

This means that the parameters of optimal models obtained by the GTLS procedure for finite observed data converge to the parameters of models that are optimal for the data generating process. A similar result can be obtained

on the behavioural level, that is, independent of the chosen parametrization. In order to formulate consistency in behavioural terms we use the so-called gap metric

$$d(\mathcal{B}_1, \mathcal{B}_2) = \|\hat{G}_{\theta_1}\hat{G}_{\theta_1}^* - \hat{G}_{\theta_2}\hat{G}_{\theta_2}^*\|_\infty \tag{23}$$

where $\|\cdot\|_\infty$ denotes the supremum norm on the unit circle, and θ_j is an isometric representation of $\mathcal{B}_j$. Note that $d(\mathcal{B}_1, \mathcal{B}_2)$ measures only the distance between the controllable parts of the systems $\mathcal{B}_i$, as $d(\mathcal{B}_1, \mathcal{B}_2) = 0$ if $\mathcal{B}_1$ and $\mathcal{B}_2$ have the same controllable part. For given $\rho < 1$ we denote by $\mathbf{B}^*(\rho)$ the set of optimal systems over Θ_ρ for the data generating process, in the sense of minimizing (6) over this set of systems, and by $\mathbf{B}_N^*(\rho)$ the set of GTLS systems minimizing (7) over Θ_ρ, so that $\mathbf{B}^*(\rho) = \{\mathcal{B}(\theta) \; ; \; \theta \in \Theta_\rho^*\}$ and $\mathbf{B}_N^*(\rho) = \{\mathcal{B}(\theta_N) \; ; \; \theta_N \in \Theta_{\rho,N}^*\}$.

Theorem 6 *For fixed complexity (m, n) and $\rho < 1$, GTLS is consistent on the behavioural level as almost surely for $N \to \infty$ there holds*

$$\sup_{\{\mathcal{B}_N \in \mathbf{B}_N^*(\rho)\}} \inf_{\{\mathcal{B} \in \mathbf{B}^*(\rho)\}} d(\mathcal{B}_N, \mathcal{B}) \to 0.$$

So the GTLS systems identified from observed data converge to optimal systems for the data generating process if the number of observations tends to infinity, and this provides an answer to the consistency question formulated at the end of Section 2.

4.2   Existence and Uniqueness of Optimal Models

In order to give these consistency results more meaning we investigate the existence and uniqueness of optimal systems. The optima are unique if $\mathbf{B}^*(\rho)$ and $\mathbf{B}_N^*(\rho)$ are singletons, and in this case we call the minima of $e(\theta)$ and $e_N(\theta)$ unique up to equivalence. That is, if (A, B, C, D) are minimal and optimal parameters then all other optimal parameters are of the form (UAU', UBV, CU', DV) with U and V unitary matrices. Of course, because of the involved nonlinearities in $e(\theta)$ and $e_N(\theta)$ minima need not be unique (up to equivalence). Minima need not even exist for $e_N(\theta)$, and for a counter example we refer to Section 5.1 in Roorda (1995a). Such non-existence of minima only occurs in exceptional cases, as is made more explicit

in the next result. Here we use the following concept of genericity of data generating processes, in terms of the spectrum $\Sigma(e^{i\omega})$ of the process w. Let $\mathcal{S}$ be the set of $q \times q$ spectral density matrices that are bounded on the unit circle, and let a metric on $\mathcal{S}$ be defined by $d(\Sigma_1, \Sigma_2) = \|\Sigma_1 - \Sigma_2\|_\infty :=$ $\sup_{\omega \in [-\pi,\pi]} \lambda_{max}\{\Sigma_1(e^{i\omega}) - \Sigma_2(e^{i\omega})\}$. Then a subset $\mathcal{S}' \subset \mathcal{S}$ is called generic if it contains a subset $\mathcal{S}'' \subset \mathcal{S}'$ that is open and dense in $\mathcal{S}$. Further we define the diameter of a set of systems $\mathbf{B}$ as the supremum of the distance $d(\mathcal{B}_1, \mathcal{B}_2)$ with $\mathcal{B}_1, \mathcal{B}_2 \in \mathbf{B}$. We consider the following properties, that of course depend on the data generating process (DGP).

- P1 : there exists an optimal system.

- P2 : the set of optimal systems has diameter at most δ.

- P3 : all optimal systems are controllable and have full complexity (m, n).

- P4 : the optimal system is unique.

Theorem 7   *(i) For every $\rho \leq 1$, the criterion e satisfies P1 for all DGP, P2 (for every fixed $\delta > 0$) and P3 for generic DGP, and P4 for a dense set of DGP.*

  (ii) For every $\rho < 1$ and for generic DGP, the criterion e_N satisfies P1, P2 (for every fixed $\delta > 0$) and P3 for N sufficiently large, almost surely.

Loosely speaking this means that, up to arbitrary finite precision, optimal systems are generically unique and the same holds true for the GTLS system for sufficiently large sample size.

4.3   Identification by Zero Extensions

The proof of consistency in Theorems 5 and 6 is based on Lemma 3(i) so that it is crucial that the stability bound $\rho < 1$ is given *a priori*, as for $\rho = 1$ uniform stability no longer holds true. We remark that generically, because of Theorem 7(i), optimal systems for the process are minimal so that $\rho < 1$, and in practice it is perhaps not a very strong restriction to fix ρ close to unity. However, if information on the stability bound is not available and one would not like to impose this condition artificially, the above result is of little

value. Therefore we now consider modelling without this prior information. As the analysis of the GTLS procedure turns out to be relatively complicated in this case, we consider the slightly adjusted criterion function (21). This is the GTLS criterion when applied to the observed data extended with zeros in past and future. Stated in terms of the algorithm (15) - (19) this means the following. We no longer optimize over the end state $x(N + 1)$ in (15), as this state is now zero, and further the errors of the approximation in (19), or equivalently in (17), over time instants $t \leq 0$ are also taken into account. Of course, minimization of (21) will keep the additional errors small as the observations are zero outside the time interval $[1, N]$, but as compared with GTLS (7) there will be some loss of fit. This additional error will be relatively smaller for systems with stronger stability bound, as the value of $\rho = \lambda_{max}(A)$ measures the rate at which the influence of wrongly specified initial conditions in (15) - (19) dies out.

For simplicity we assume that the data generating process is Gaussian ARMA, that is, in the Wold representation (5) ε is Gaussian white noise and T is a rational transfer function. These assumptions can be weakened, for example to those stated in Brillinger (1975, Theorem 7.7.2). In order to prove consistency we first state an auxiliary result.

Lemma 8 *Under the above assumptions there holds that*

(i) $e_N^0 \to e$ uniformly over Θ if $N \to \infty$, almost surely;

(ii) the global minimum of e_N^0 over Θ always exists.

We use the following notation. By $\Theta^* = \text{argmin}_{\{\theta \in \Theta\}} e(\theta)$ we denote the set of optimal parameters for the process and by $\Theta_N^{0*} = \text{argmin}_{\{\theta \in \Theta\}} e_N^0(\theta)$ the set of parameters that minimize the adjusted GTLS criterion (21). These sets are non-empty because of Theorem 7(i) and Lemma 8(ii). For simplicity, and motivated by the results in Theorem 7(i) we will assume that Θ^* parametrizes a single controllable system $\mathcal{B}^*$ of full complexity. Further we denote by $\mathbf{B}_N^{0*}$ the set of systems parametrized by Θ_N^{0*}. The next theorem states consistency results on the parametric and on the behavioural level.

Theorem 9 *Under the above assumptions,*

(i) *identification by adjusted GTLS (21) is consistent over the full param-
     eter set Θ, as for $N \to \infty$ there holds that almost surely*

$$\sup_{\{\theta_N^0 \in \Theta_N^{0*}\}} \inf_{\{\theta \in \Theta^*\}} \|\theta_N^0 - \theta\| \to 0.$$

(ii) *Consistency holds also true on the behavioural level, as $\mathbf{B}_N^{0*}$ is non-
      empty and for $N \to \infty$ there holds almost surely*

$$\sup_{\{\mathcal{B} \in \mathbf{B}_N^{0*}\}} d(\mathcal{B}, \mathcal{B}^*) \to 0.$$

This shows that identification by applying GTLS to the data extended with
zero values in past and future gives consistent estimates over the full param-
eter set Θ, that is, over the class of all linear systems with complexity at
most (m, n).

5   Simulation Examples

We illustrate the foregoing consistency results by means of two simple sim-
ulations. Here we will apply the GTLS procedure on the full parameter set
Θ, that is, without conditions on the stability of the system. We also con-
sider the effect of extending the data by zeros in past and future, that is,
identification with the criterion (21) instead of (7). In both examples we
first determine the optimal approximation on the process level, that is, the
system of restricted complexity that minimizes (6). Then we apply GTLS
to data generated by the process, and we investigate whether the identified
system converges to the optimal approximation if the number of observations
increases. The GTLS algorithm (15) - (19) is implemented by the procedures
described in Roorda (1995a, Appendix B). [1]

5.1   Mexican Hat

The so-called Mexican hat is a filter for change detection in noisy signals,
where the output is obtained as the second derivative of the input signal after
smoothing with the normal density $h(s) = (1/\sqrt{2\pi})e^{-s^2/2}$. Let the input

---

[1]The authors thank Berend Roorda for allowing us to use his Matlab procedures.

be denoted by u and the output by y, then these variables are related by $y(t) = -\frac{d^2}{dt^2}\{\int_{-\infty}^{\infty} h(s)u(t-s)ds\}$. In this simulation we consider a discretized and scaled version $y(t) = \sum_{k=-50}^{50} H_k u(t-k)$ where $H_k = -10h''(k/10)$ with $h''(s) := -\frac{d^2}{ds^2}h(s)$. Note that this is not a causal system, as the output depends on future values of the input. The data generating process is given by $w = (w_1, w_2)'$ with $w_1 = u+e_1$ and $w_2 = y+e_2$, where y is generated from u by the Mexican hat filter and where (u, e_1, e_2) is a Gaussian white noise process with independent components, mean zero, and standard deviations respectively $\sigma_u = 1$, $\sigma_1 = 0.05$ and $\sigma_2 = 0.73$. This has been chosen so that the signal-to-noise ratio is twenty for both variables. The data generating process corresponds to a process of type (3), with latent process $(u, y)'$ and noise process $(e_1, e_2)'$. The Mexican hat is not a rational transfer function, and our discrete time version has McMillan degree 100 so that the order of the polynomial equation (4) for the latent variables $(u, y)'$ has order 100. The idea of this simulation example is to approximate the data generating process by a linear system with lower complexity. We will fix the complexity at $(m, n) = (1, 4)$, and we mention that values of the state dimension $n \geq 6$ give very accurate approximations.

First we characterize the optimal approximation of the process, that is, the system with complexity $(m, n) = (1, 4)$ that minimizes (6). As stated in Section 3.3 this is given by the GTLS model for the standardized Wold transfer function T in (5). However, in this example it is simpler to apply GTLS to another transfer function that generates the process from standard white noise, that is, $w = T_0\varepsilon_0$ where $\varepsilon_0 = (u, e_1/\sigma_1, e_2/\sigma_2)$ is a white noise process with zero mean and unit covariance matrix and where

$$T_0 = \begin{pmatrix} 1 & \sigma_1 & 0 \\ H & 0 & \sigma_2 \end{pmatrix}.$$

Here $H = \sum_{k=-50}^{50} H_k z^{-k}$ is the discrete time version of the Mexican hat. The optimal approximation is given by the linear system that minimizes the sum of the squares of the GTLS errors of the three columns in T_0. In this case, because the coefficients of the filter T_0 are zero outside a range of length 101, the optimal approximation can simply be calculated by concatenation, that is, we consider an 'observation' length 101 for each of the columns of T_0 and apply GTLS to the resulting 2×303 data matrix. The GTLS model has an error (6) of approximately 0.693, and the coefficients of the polynomial

relation function (4) are given in Table 1. This relation can be transformed to a (non-causal) transfer function H^a from w_1 to w_2, and the resulting relative error $\|H^a - H\|/\|H\|$ is also given in Table 1 where $\|H\| = (\sum_k H_k^2)^{1/2}$.

We generated data on observation intervals of lengths N =50, 100, 200, 500 and 1000 and identified GTLS systems of complexity $(m, n) = (1, 4)$. Of course these systems are random, and we will report the result of an arbitrary simulation. Other simulations gave similar results, but we will not discuss further details of this random variation. The results in terms of the coefficients of relation (4) and the relative errors $\|H^N - H\|/\|H\|$ and $\|H^N - H^a\|/\|H^a\|$ of the identified transfer function H^N are in Table 1. This indicates consistency. It requires relatively large sample sizes to come close to the optimal approximation H^a as it is relatively difficult to identify the zeros of this system. The poles are very well identified, notwithstanding the fact that H^a has two stable poles and two unstable ones.

N	50	100	200	500	1000	∞
$\|H^N - H\|/\|H\|$	1.086	0.503	0.340	0.318	0.302	0.288
$\|H^N - H^a\|/\|H^a\|$	1.124	0.500	0.151	0.059	0.042	0.000
$w_1(t)$	0.353	0.061	0.261	0.175	0.171	0.086
$w_1(t-1)$	0.078	0.158	0.453	0.394	0.362	0.141
$w_1(t-2)$	0.046	0.225	0.411	0.508	0.453	0.156
$w_1(t-3)$	0.253	0.361	0.280	0.417	0.383	0.133
$w_1(t-4)$	0.305	0.305	0.139	0.185	0.181	0.081
$w_2(t)$	1.000	1.000	1.000	1.000	1.000	1.000
$w_2(t-1)$	-1.561	-3.616	-3.994	-3.947	-3.941	-3.940
$w_2(t-2)$	1.251	4.948	6.050	5.907	5.890	5.885
$w_2(t-3)$	-2.054	-3.042	-4.112	-3.970	-3.952	-3.946
$w_2(t-4)$	1.395	0.713	1.059	1.011	1.006	1.003

Table 1: Results of Example 1. The first row shows the sample size, and the next two rows give the relative errors of H^N with respect to the Mexican hat H and the optimal approximation H^a respectively. The next rows show the GTLS coefficients of the relation (4) in the order $w_1(t), \ldots, w_1(t-4), w_2(t), \ldots, w_2(t-4)$ and scaled so that $w_2(t)$ has coefficient one. The columns correspond to the different sample sizes and the last column shows the optimal process approximation.

	N	10	50	100	500	∞
GTLS	$w_1(t)$	0.049	0.019	0.012	0.000	0.000
	$w_1(t-1)$	0.991	1.000	0.999	1.000	1.000
	$w_2(t)$	1.000	1.000	1.000	1.000	1.000
	$w_2(t-1)$	0.104	0.012	0.002	0.004	0.000
aGTLS	$w_1(t)$	0.126	0.003	0.010	0.000	0.000
	$w_1(t-1)$	0.955	0.976	0.998	1.000	1.000
	$w_2(t)$	1.000	1.000	1.000	1.000	1.000
	$w_2(t-1)$	0.207	0.034	0.003	0.006	0.000

Table 2: Results of Example 2. The first row shows the sample size. The next four rows give the GTLS coefficients of the relation (4) in the order $w_1(t), w_1(t-1), w_2(t), w_2(t-1)$ and scaled so that $w_2(t)$ has coefficient one. The last four rows are similar for the adjusted GTLS criterion (21) instead of (7). The columns correspond to the different sample sizes and the last column shows the optimal process approximation.

5.2 Bivariate Autoregression

In the second simulation we consider a simple bivariate autoregressive process. The data are generated by $w = (w_1, w_2)'$ with $w_1(t) = 0.9w_1(t-4) + \varepsilon_1(t)$ and $w_2(t) = w_1(t-1) + \varepsilon_2(t)$, where $\varepsilon = (\varepsilon_1, \varepsilon_2)'$ is a Gaussian white noise process with zero mean and unit covariance matrix. We consider approximations by linear systems of complexity $(m, n) = (1, 1)$. The optimal relation function for this process is computed from the Wold representation (5) as discussed in Section 3.3. This gives the approximation $\hat{w}_2(t) = \hat{w}_1(t-1)$ with error (6) equal to $\frac{1}{2}\sqrt{2}$.

We generated data on observation intervals of lengths $N =$ 10, 50, 100 and 500 and identified GTLS models of complexity $(m, n) = (1, 1)$. Further we also identified models by the adjusted GTLS criterion (21). The results are in Table 2 in terms of the estimated polynomial equations (4). This again confirms the consistency of GTLS identification. In addition, it indicates that the criterion (21) corresponding to extending the data with zeros in past and future is close to the GTLS criterion (7) if the number of observed data is sufficiently large.

6   Conclusion

In this paper we presented a statistical analysis for the identification of open symmetric models. This means that all variables are treated in a similar way and that the model leaves some aspects of the evolution of the variables unexplained. This is in contrast with more conventional approaches in system identification, where models are usually either closed and symmetric, such as ARMA, or open and asymmetric, such as ARMAX. In our approach the object of interest is the system behaviour, and basic system properties are expressed independent of parametrization. We analysed the global total least squares (GTLS) method for linear system identification within a stochastic framework. Here the complexity of the models is fixed, and a final model choice will involve a trade-off between complexity and goodness of fit.

The central result concerns the consistency of GTLS, in the sense that the identified model converges to an optimal approximation of the data generating process. In terms of isometric state representations, it is shown that GTLS is consistent if the maximum modulus of the eigenvalues of the state transition matrix is bounded a priori by a fixed number $\rho < 1$. The full parameter set has $\rho \leq 1$ and, in fact, representations with $\rho = 1$ are not minimal and hence correspond to systems of lower complexity. Consistency for the whole model class is obtained if the observed data are extended by zeros in past and future, and GTLS is applied to these extended data. For finite data this may cause an increase in error as compared to GTLS, but this effect vanishes asymptotically with a speed depending on ρ.

The question whether GTLS is consistent on the whole model class is a topic for future investigation. A further statistical analysis of identification by means of open symmetric models is needed. In particular, results on (asymptotic) distributions would be of interest, both in terms of parameters and in terms of system behaviours. This would open the way to testing procedures, for instance concerning the selection of model complexity.

7   Appendix

Proof of Proposition 1
For (i) and (ii) we refer to Roorda (1995a), Propositions 3.2.4, 3.2.7 and 4.2.2, (iii) follows directly from Heij et al. (1995), Theorem 5, and (iv) for minimal θ is

stated as Algorithm 1 in Roorda (1995a), see also Roorda(1995b). The result in (iv) for non-minimal θ follows because in this case the minimization for e_N should be restricted to the set of reachable states $x(N+1)$, while for e_N^* this restriction is not taken into account. $\square$

Proof of Proposition 2

(i) It is evident from (12) that Θ is a closed and bounded subset of $\mathsf{R}^{(n+q)(n+m)}$, so that it is compact. As Θ_ρ in (20) is a closed subset of Θ it is also compact.

(ii) Rewrite (14) as $e^2(\theta) = \int_{-\pi}^{\pi} \mathrm{trace}\{(I - \hat{G}_\theta \hat{G}_\theta^*)\Sigma\}d\omega$, where $\hat{G}_\theta(e^{i\omega}) = D + \sum_{k=1}^{\infty} CA^{k-1}Be^{i\omega k}$. Let $\{\theta_k\}$ be a sequence in Θ that converges to $\theta_0 = (A_0, B_0, C_0, D_0)$ for $k \to \infty$, then it suffices to prove that $e^2(\theta_k) \to e^2(\theta_0)$. As the standardized Wold representation (5) is assumed to be absolutely summable, the spectral density Σ is bounded on the unit circle so that it suffices to prove that $\int \mathrm{trace}\{\hat{G}_k \hat{G}_k^* - \hat{G}_0 \hat{G}_0^*\}d\omega \to 0$, where $\hat{G}_k := \hat{G}_{\theta_k}$. This is a standard result in case A_0 has all its eigenvalues within the open unit disc, as $\hat{G}_0$ is then a bounded rational matrix function on the unit circle. Otherwise, let $\{e^{i\omega_l}; l = 1, \ldots, L\}$ be the eigenvalues of A_0 on the unit circle, and define for $\delta > 0$ the sets $U_1 := \bigcup_{l=1}^{L}\{\omega \in [-\pi, \pi]; |\omega - \omega_l| < \delta\}$ and $U_2 := [-\pi, \pi] \backslash U_1$. Then $\int_{U_2} \mathrm{trace}\{\hat{G}_k \hat{G}_k^* - \hat{G}_0 \hat{G}_0^*\}d\omega \to 0$ as before, and $|\int_{U_1} \mathrm{trace}\{\hat{G}_k \hat{G}_k^* - \hat{G}_0 \hat{G}_0^*\}d\omega| \leq \int_{U_1} \mathrm{trace}\{\hat{G}_k \hat{G}_k^* + \hat{G}_0 \hat{G}_0^*\}d\omega = \int_{U_1} \mathrm{trace}\{\hat{G}_k^* \hat{G}_k + \hat{G}_0^* \hat{G}_0\}d\omega = 2m\int_{U_1} d\omega = 4mL\delta \leq 4mn\delta$. Letting $\delta \downarrow 0$ the result follows.

(iii) From the definition in (7) it follows that the GTLS error is obtained by the residual after projecting the observed data w_N onto the linear space $\mathcal{B}_{|[1,N]}$. If $\mathcal{B}$ has complexity (m, n) then for $N \geq n$ the linear space $\mathcal{B}_{|[1,N]} \subset \mathsf{R}^{Nq}$ has dimension $n + Nm$. The representation (18) shows that this projection space can be represented as

$$\hat{W}_N = F_N \hat{x}(1) + G_N \hat{V}_N \qquad (24)$$

where $\hat{W}_N := (\hat{w}(1)', \ldots, \hat{w}(N)')'$, $\hat{V}_N := (\hat{v}(1)', \ldots, \hat{v}(N)')'$, $F_N := (C', A'C', \ldots, (A')^{N-1}C')'$, and G_N is the $Nq \times Nm$ matrix with t-th block row, $t = 1, \ldots, N$, given by $(CA^{t-2}B, \ldots, CAB, CB, D, 0, \ldots, 0)$. In order to obtain time series in $\mathcal{B}_{|[1,N]}$, $\hat{V}_N$ can be chosen freely and $\hat{x}(1)$ should be reachable in (18). If θ is minimal then all states $\hat{x}(1)$ are reachable as in this case $[A \ B]$ has full row rank, and then the matrix $H(\theta) := [F_N, \ G_N]$ has full column rank $Nm + n$.

Now let be given observations w_N and minimal $\theta_0 \in \Theta$, and let $\theta_k \to \theta_0$ for $k \to \infty$, so that θ_k is also minimal for k sufficiently large. According to the foregoing, $e_N(\theta_0)$ is obtained by projecting w_N onto $\mathcal{B}(\theta_0)_{|[1,N]} = \mathrm{im}\, H(\theta_0)$, and $e_N(\theta_k)$ is obtained by projecting onto $\mathrm{im}\, H(\theta_k)$. So the continuity of e_N is equivalent to

the continuity of these projections as function of θ. As the entries of $H(\theta)$ are clearly continuous functions of θ, continuity of the projections is equivalent to the condition that the projection spaces have constant dimension. Because θ_0 and θ_k are minimal it follows that rank$H(\theta_0) = Nm + n = \text{rank}H(\theta_0)$ for k sufficiently large.

(iv) In a first step we construct the set of reachable states $\mathcal{X} \subseteq \mathsf{R}^n$ of a system $\mathcal{B}(\theta)$: Let the matrices X_j be defined by the recursion $X_0 = I$ and $X_{j+1} = [AX_j, B]$. Then by an induction argument it follows that $\text{im } X_{j+1} \subseteq \text{im } X_j$ and if $\text{im } X_{l+1} = \text{im } X_l$ for some l, then $\text{im } X_{l+j} = \text{im } X_l$ holds for all $j \geq 0$. This immediately implies that $\text{im } X_{n+1} = \text{im } X_n$. Clearly $\text{im } X_n \supseteq \mathcal{X}$ holds by construction and $\text{im } X_n \subseteq \mathcal{X}$ follows from $\text{im } X_{n+1} = \text{im } X_n$. Thus we have proved that the set of reachable states equals the image of the matrix X_n.

Therefore $e_N(\theta)$ may be obtained by projecting W_N onto the column space of the matrix $\bar{H}(\theta) = [F_N(\theta)X_n(\theta), G_N(\theta)]$. Now we construct a unitary matrix $U = [U_1, U_2]$, such that $\bar{H}(\theta_0)U_1$ has full column rank and $\bar{H}(\theta_0)U_2 = 0$. Let $\bar{e}(\theta)$ denote the error of the projection of W_N onto the column space of $\bar{H}(\theta)U_1$. Clearly $e_N(\theta) \leq \bar{e}(\theta)$ and $e_N(\theta_0) = \bar{e}(\theta_0)$ holds. Finally, since $\bar{H}(\theta_k)U_1$ converges to the full rank matrix $\bar{H}(\theta_0)U_1$ we have $\limsup_k e_N(\theta_k) \leq \lim_k \bar{e}(\theta_k) = \bar{e}(\theta_0) = e_N(\theta_0)$.
$\square$

For later reference we collect some results in the following lemma.

Lemma 10   *(i) Let $\{G_\theta; \theta \in \Theta\}$ and $\{F_\theta; \theta \in \Theta\}$ be two uniformly stable classes of filters, then $\{G_\theta F_\theta; \theta \in \Theta\}$ is also uniformily stable.*

(ii) Let ε be a white noise process satisfying the assumptions stated at the end of Section 2.1 and let $\{G_\theta; \theta \in \Theta\}$ be uniformily stable. Then for $w_\theta = G_\theta \varepsilon$ there holds that almost surely $\sup_\theta \|\frac{1}{N} \sum_{t=1}^N w_\theta(t)w_\theta'(t) - \mathrm{E}\, w_\theta(t)w_\theta'(t)\| \to 0$.

Proof. Part (i) is immediate and (ii) is a simple generalization of Ljung (1987, Theorem 2.B.1 and Corollary to Theorem 2.B.1) to the case of non-causal filters.
$\square$

Proof of Lemma 3

(i) As Θ_ρ is compact, the corresponding set of $n \times n$ matrices A is also compact. For $\lambda_{max}(A) \leq \rho < 1$ it follows from Davis and Vinter (1985, Proposition D.3.1) that for every $\rho < \lambda < 1$ there exists $c > 0$ such that $\|A^k\| \leq c\lambda^k$ for all $k = 0, 1, 2, \ldots$, uniformly over Θ_ρ. Here $\|\cdot\|$ is any matrix norm, and we consider the induced norm. For the filters $\tilde{G}_\theta(z^{-1}) = \sum_{k=0}^\infty \tilde{G}_\theta(k)z^{-k}$ it follows from this result and (13) that $\|\tilde{G}_\theta(0)\| = \|\tilde{D}\| \leq 1$ and $\|\tilde{G}_\theta(k)\| = \|CA^{k-1}\tilde{B}\| \leq \|C\|\|A^{k-1}\|\|\tilde{B}\| \leq c\lambda^{k-1}$ for $k \geq 1$, that is, there is a constant $M > 0$ so that $\|\tilde{G}_\theta(k)\| \leq M\lambda^k$ for all

$k \geq 0$. So $\sum_k \sup_\theta \|\tilde{G}_\theta(k)\| \leq M \sum \lambda^k < \infty$ which proves the uniform stability of $\tilde{G}_\theta^*$ and, by the above Lemma 10(i), of P_θ.

For $\rho = 1$ the family of filters is not uniformly stable (unless in the trivial cases $n = 0$ and $m = q$). We prove this by construction for $(q, m, n) = (2, 1, 1)$, for higher dimensions the result follows by taking this construction as a subsystem. Let $(A, \tilde{B}, C, \tilde{D}) = (\alpha, -\gamma\beta, \gamma(1, 1)', \alpha\beta(0, 1)')$ with $0 < \alpha < 1$ and β and γ so that $(A, \tilde{B}, C, \tilde{D})$ is an isometric representation, that is, $\gamma = \sqrt{(1 - \alpha^2)/2}$ and $\beta = \sqrt{2/(1 + \alpha^2)}$. These parameters are minimal, and $\|\tilde{G}_\theta(k)\| = \|CA^{k-1}\tilde{B}\| = (1 - \alpha^2)\alpha^{k-1}/\sqrt{1 + \alpha^2} \geq (1 - \alpha^2)\alpha^{k-1}/\sqrt{2}$ for $k \geq 1$. A simple calculation shows that $\sup_\alpha (1 - \alpha^2)\alpha^{k-1}/\sqrt{2}$ is obtained for $\alpha = \sqrt{1 - 2/(k + 1)}$ with value $c_k/(k + 1)$ where $c_k = \sqrt{2}(1 - 2/(k + 1))^{(k-1)/2}$. As $c_k \to \sqrt{2}/e$ for $k \to \infty$ it follows that this family of filters is not uniformly stable.

(ii) As realizations of the process $\hat{w}_\theta$ are elements of the system $\mathcal{B}(\theta)$ it follows that $e_N^2(\theta) \leq \frac{1}{N}\tilde{W}_N'(\theta)\tilde{W}_N(\theta)$. According to (15) and (17) we have

$$\tilde{V}_N(\theta) = M_N(\theta)x(N + 1) + L_N(\theta)W_N \tag{25}$$

where $W_N = (w'(N), w'(N - 1), \ldots, w'(1))'$ are the observed data, in reversed time order, L_N is the $Nm \times Nq$ matrix with t-th block row, $t = 1, \ldots, N$, given by $(\tilde{B}'(A')^{t-2}C', \ldots, \tilde{B}'A'C', \tilde{B}'C', \tilde{D}', 0, \ldots, 0)$ and $M_N = (\tilde{B}, A\tilde{B}, \ldots, A^{N-1}\tilde{B})'$. According to Proposition 1(iv) the error $\{e_N^*(\theta)\}^2 = \frac{1}{N}\tilde{V}_N^*(\theta)'\tilde{V}_N^*(\theta)$ is computed by $\tilde{V}_N^*(\theta) = M_N(\theta)x^* + L_N(\theta)W_N = M_N(\theta)(x^* - x(N + 1)) + \tilde{V}_N(\theta)$ where x^* is choosen such that $\tilde{V}_N^*(\theta)'\tilde{V}_N^*(\theta)$ is minimal. Then x^* is obtained by projecting $\tilde{V}_N(\theta)$ onto the image of $M_N(\theta)$, so that $\tilde{V}_N^*(\theta)$ is the projection onto the orthogonal complement of this space, that is, $\tilde{V}_N^*(\theta) = (I - \Pi_N(\theta))\tilde{V}_N(\theta)$. $\square$

Proof of Lemma 4

(i) According to Proposition 1(iii) there holds $e^2(\theta) = \mathrm{E}\,\|\tilde{v}(t)\|^2 = \mathrm{E}\,\|\tilde{w}(t)\|^2$, and $\frac{1}{N}\tilde{V}_N'(\theta)\tilde{V}_N(\theta) = \frac{1}{N}\sum_{t=1}^N \tilde{v}_\theta'(t)\tilde{v}_\theta(t)$ and $\frac{1}{N}\tilde{W}_N'(\theta)\tilde{W}_N(\theta) = \frac{1}{N}\sum_{t=1}^N \tilde{w}_\theta'(t)\tilde{w}_\theta(t)$ are the corresponding sample variances. So we have to prove that the sample variances converge almost surely to the process variances, uniformly for the class of filters given by $\tilde{v} = \tilde{G}_\theta w = \tilde{G}_\theta T\varepsilon$, and $\tilde{w} = P_\theta w = P_\theta T\varepsilon$. This follows by the uniform stability result in Lemma 3(i) and by Lemma 10(ii).

(ii) We use the notation introduced in the proof of Lemma 3. The projection operator Π_N is given in terms of the matrix $M_N = (\tilde{B}, A\tilde{B}, \ldots, A^{N-1}\tilde{B})'$ by $\Pi_N = M_N(M_N'M_N)^- M_N' = M_N K_N^2 M_N'$, where $(M_N'M_N)^-$ denotes the positive semidefinite generalized inverse and K_N is its positive semidefinite square root. We remark that the pair $(A, \tilde{B})$ need not be controllable. Now define the family of filters $H_{\theta,k}(z^{-1}) = \sum_{j=0}^\infty H_{\theta,k}(j)z^{-j}$, with $\theta \in \Theta_\rho, k \in \mathsf{N}$, by $H_{\theta,k}(j) = K_k A^j \tilde{B}$ for $j \leq k$ and $H_{\theta,k}(j) = 0$ for $j > k$, and define corresponding processes by $u_{\theta,k} = $

$H_{\theta,k}\tilde{v} = H_{\theta,k}\tilde{G}_\theta^* T\varepsilon$. Then by construction there holds $\frac{1}{N}\tilde{V}_N'(\theta)\Pi_N(\theta)\tilde{V}_N(\theta) = \frac{1}{N}\|u_{\theta,N}(N)\|^2$ and we have to prove that almost surely

$$\sup_{\{\theta\in\Theta_\rho\}} \frac{1}{N}\|u_{\theta,N}(N)\|^2 \to 0. \tag{26}$$

Now assume that $\{H_{\theta,k}; \theta\in\Theta_\rho, k\in\mathsf{N}\}$ is a uniformly stable family of filters. As T is bounded and $\{\tilde{G}_\theta^*; \theta\in\Theta_\rho\}$ is uniformly bounded, it then follows from Lemma 10 that almost surely and uniformly over $\Theta_\rho\times\mathsf{N}$ there holds $\frac{1}{N}\|\sum_{t=1}^N u_{\theta,k}(t)u_{\theta,k}'(t) - \mathrm{E}(u_{\theta,k}(t)u_{\theta,k}'(t))\| \to 0$ and hence also $\frac{1}{N}[\sum_{t=1}^N[\|u_{\theta,k}(t)\|^2 - \mathrm{E}\|u_{\theta,k}(t)\|^2] \to 0$. Applying this result for N and $N-1$ shows that $\frac{1}{N}[\|u_{\theta,N}(N)\|^2 - \mathrm{E}\|u_{\theta,N}(N)\|^2] \to 0$ uniformly over Θ_ρ. Because of uniform stability, the variance of $u_{\theta,N}$ is uniformily bounded, so that (26) follows.

It remains to prove the uniform stability of $\{H_{\theta,k}; \theta\in\Theta_\rho, k\in\mathsf{N}\}$. First we prove an auxiliary result for the term A^j in this expression. The Cayley-Hamilton theorem states that $A^n = \sum_{k=0}^{n-1}\alpha_k A^k$ where $\det(zI - A) = z^n - \sum_{k=0}^{n-1}\alpha_k z^k$ is the characteristic polynomial. For all $j\geq n$ we can write $A^j = \sum_{k=0}^{n-1}\alpha_k(j)A^k$, and it is easily checked that the coefficient vector $\alpha(j) = (\alpha_0(j), \alpha_1(j),\ldots,\alpha_{n-1}(j))'$ can be recursively computed by $\alpha(j+1) = \Omega\alpha(j)$ with $\alpha(n) = \alpha$ and with Ω the $n\times n$ companion matrix with last column α and first $n-1$ columns $\begin{pmatrix} 0 \\ I_{n-1} \end{pmatrix}$. As the set of A- matrices in Θ_ρ is compact and α is a continuous function of A it follows that $\|\alpha\|$ is bounded. The set of Ω-matrices over Θ_ρ is also compact, and as A and Ω have the same eigenvalues it also follows that $\lambda_{max}(\Omega)\leq\rho<1$. It follows again by Davis and Vinter (1985, Proposition D.3.1) that for every $\rho<\lambda<1$ there exists $c'>0$ such that $\|\Omega^k\|\leq c'\lambda^k$ uniformly over Θ_ρ. Together with the boundedness of $\|\alpha\|$ this shows that there is a constant $M'>0$ such that $\|\alpha(j)\|\leq M'\lambda^j$. Further for $i\leq n\leq k$ there holds $M_k'M_k\geq M_n'M_n\geq A^{i-1}\tilde{B}\tilde{B}'(A')^{i-1}$ so that $K_k\leq K_n$ and $\|K_n A^i\tilde{B}\|\leq 1$. So for $j,k\geq n$ there holds $\|K_k A^j\tilde{B}\|\leq\|K_n A^j\tilde{B}\|\leq \sum_{i=0}^{n-1}|\alpha_i(j)|\cdot\|K_n A^i\tilde{B}\|\leq nM'\lambda^j$, and this shows that the class of filters $H_{\theta,k}$ is uniformly stable over $\Theta_\rho\times\mathsf{N}$.

(iii) This is immediate from (i), (ii) and (22). $\square$

Proof of Theorem 5
We prove this by contradiction. So suppose that there would be a sequence $\theta_N\in\Theta_{\rho,N}^*$ and $\delta>0$ such that $\inf_{\{\theta\in\Theta_\rho^*\}}\|\theta_N - \theta\|\geq\delta$. As Θ_ρ is compact, see Proposition 2(i), there is a subsequence θ_{N_k} and $\theta_0\in\Theta_\rho$ so that $\|\theta_{N_k} - \theta_0\|\to 0$ for $k\to\infty$. For simplicity of notation we relabel this subsequence as θ_N. Because of the results in Proposition 2(ii) and Lemma 4(iii) it follows that almost surely $|e_N(\theta_N) - e(\theta_0)|\leq |e_N(\theta_N) - e(\theta_N)| + |e(\theta_N) - e(\theta_0)|\to 0$ for $N\to\infty$. As

$\theta_N \in \Theta^*_{\rho,N}$ it follows that $e_N(\theta_N) \leq e_N(\theta)$ for all $\theta \in \Theta_\rho$. Combining these two results it follows that $e(\theta_0) = \lim_N e_N(\theta_N) \leq \lim_N e_N(\theta) = e(\theta)$ for all $\theta \in \Theta_\rho$, so that $\theta_0 \in \Theta^*_\rho$. This means that $\inf_{\{\theta \in \Theta^*_\rho\}} \|\theta_N - \theta\| \leq \|\theta_N - \theta_0\| \to 0$, which contradicts the assumption that this infimum was at least δ. This proves the result. $\square$

Proof of Theorem 6
According to the result in Theorem 5 it suffices to prove that $d(\mathcal{B}(\theta), \mathcal{B}(\theta_0)) \to 0$ if $\|\theta - \theta_0\| \to 0$, that is, that $\mathcal{B}(\theta)$ is continuous on Θ_ρ for $\rho < 1$. This is a standard result, as the functions $\hat{G}_\theta \hat{G}^*_\theta$ in (23) are continuous on Θ_ρ. The crucial point to notice here is that the function $(zI - A)^{-1}$ is uniformly bounded on the unit circle because $\lambda_{max}(A) \leq \rho < 1$. We mention that continuity does not hold true on Θ, see Heij et al. (1995, Proposition 11(iii)). $\square$

Proof of Theorem 7 (i) Throughout we take (m, n) and $\rho \leq 1$ fixed.

Property P1 follows from Proposition 2 (i) and (ii).

To prove P3 we first state an auxiliary lemma that we will prove later.

Lemma 11 *Every DGP with rational spectrum Σ, such that each of the entries σ_{ij} has at least $2n$ poles within the unit circle that are not poles of the other entries σ_{kl}, $(k,l) \neq (i,j)$, satisfies P3.*

Because every stationary process can be approximated arbitrarily well by DGP's that satisfy the above conditions, it follows from this lemma that P3 is a dense property. Let Σ_0 satisfy P3. Because the set Θ' of parameters θ that are non-minimal or for which $\mathcal{B}(\theta)$ is non-controllable form a closed and hence compact subset of Θ_ρ, it follows from Proposition 2(ii) that for the DGP Σ_0 the minimum of the errors of systems $\mathcal{B}(\theta)$ with $\theta \in \Theta'$ is strictly larger than the minimum over Θ_ρ. Because the error is a continuous function of the spectrum Σ, it follows that the same holds true in an open neighbourhood of Σ_0. So P3 is dense and open, that is, generic.

Next we show that P4 holds for a dense set of DGP. Let w_0 be a given DGP satisfying P3 and let $\mathcal{B}_0$ be an optimal controllable system of complexity (m, n) over Θ_ρ, so that $e(\mathcal{B}_0) \leq e(\mathcal{B}(\theta))$ for all $\theta \in \Theta_\rho$. Let $\tilde{w}_0 = (I - P_0)w_0$ where P_0 is the operator of orthogonal projection onto $\mathcal{B}_0$, and for $\delta > 0$ define the process $w_\delta := w_0 - \delta \tilde{w}_0$. Using the notation $\|w\| = \{E\|w(t)\|^2\}^{1/2}$, the error of the system $\mathcal{B}$ with projection P for the DGP w_δ is given by $e_\delta(\mathcal{B}) = \|(I - P)w_\delta\| = \|(I - P)w_0 - \delta(I - P)(I - P_0)w_0\| \geq \|(I - P)w_0\| - \delta\|(I - P)(I - P_0)w_0)\| \geq (1 - \delta)\|(I - P_0)w_0\| = (1 - \delta)e(\mathcal{B}_0) = e_\delta(\mathcal{B}_0)$. This shows $\mathcal{B}_0$ is also an optimal system for w_δ, and that another system $\mathcal{B}$ is optimal for w_δ if and only if it satisfies the two conditions $\|(I - P)w_0\| = \|(I - P_0)w_0\|$ and $\|(I - P)(I - P_0)w_0)\| = \|(I - P_0)w_0\|$.

Because $\|(I - P_0)w_0\|^2 = \|P(I - P_0)w_0\|^2 + \|(I - P)(I - P_0)w_0\|^2$, the second condition implies that $\|P(I - P_0)w_0\| = 0$, and since w_0 is a full rank process this means that $P = PP_0$ so that $P_0 - P$ is a projection operator that is orthogonal to $I - P_0$. The first optimality condition $\|(I - P)w_0\| = \|(I - P_0)w_0\|$ then implies that $\|(I - P_0)w_0\|^2 = \|(I - P)w_0\|^2 = \|(I - P_0)w_0\|^2 + \|(P_0 - P)w_0\|^2$, so that $P = P_0$ because w_0 is a full rank process. Because w_0 satisfies P3, the same holds true for w_δ for δ sufficiently small. This means that all optimal systems for w_δ are controllable and that they have the same projection operator $P = P_0$. But this means that $\mathcal{B} = \mathcal{B}_0$ for δ sufficiently small, that is, then P4 holds for w_δ. Since P3 is a dense property, this proves that P4 is a dense property.

To show P2, let Σ_0 belong to the dense set of DGP satisfying P4 with unique optimal system $\mathcal{B}_0$. Let $\delta > 0$ and define $\mathbf{B}$ as the set of systems with $d(\mathcal{B}, \mathcal{B}_0) \geq \delta/2$. As this is a closed set, it follows that for the DGP Σ_0 the minimum of the errors over $\mathbf{B}$ is strictly larger than the minimum over all systems of complexity (m, n). Because the error is a continuous function of the spectrum, it follows that for all DGP in a neighbourhood of Σ_0 the minimum error will also only be obtained for systems with $d(\mathcal{B}, \mathcal{B}_0) < \delta/2$. As this is a set of diameter at most δ, this proves that P2 is dense and open, that is, generic.

It remains to show Lemma 11. We prove this in four steps.

Step 1. For given DGP with spectrum Σ let $\mathcal{B}(\theta)$ be an optimal system in Θ_ρ and let $(\hat{G}, \tilde{G})$ be the corresponding pair of isometric transfer functions. Then

$$\int_{-\pi}^{\pi} \tilde{G}^*(e^{i\omega})\Sigma(e^{i\omega})\hat{G}(e^{i\omega})d\omega = 0.$$

This is proved as follows. Let $\tilde{G}_\epsilon = (\tilde{G}, \hat{G})(I, \epsilon X')'Y$ with $X \in \mathsf{R}^{m \times (q-m)}$ and Y such that $\tilde{G}_\epsilon^* \tilde{G}_\epsilon = Y'(I + \epsilon^2 X'X)Y = I$. As $\tilde{G}$ and $\hat{G}$ have a common denominator it follows that $\tilde{G}_\epsilon$ corresponds to a system of complexity (m, n) with representation $\theta_\epsilon \in \Theta_\rho$. Further $Y = I + O(\epsilon^2)$ and $\tilde{G}_\epsilon = \tilde{G} + \epsilon\hat{G}X + O(\epsilon^2)$, so that $\mathrm{trace}\{\int \tilde{G}_\epsilon^*\Sigma\tilde{G}_\epsilon d\omega\} = \mathrm{trace}\{\int \tilde{G}^*\Sigma\tilde{G}d\omega\} + 2\epsilon \, \mathrm{trace}\{\int \tilde{G}^*\Sigma\hat{G}d\omega X\} + O(\epsilon^2)$. Since θ is optimal there holds $\mathrm{trace}\{\int \tilde{G}^*\Sigma\hat{G}d\omega X\} = 0$ for all X and this shows the result.

Step 2. For given DGP with spectrum Σ let $\mathcal{B}(\theta)$ be an optimal system in Θ_ρ and assume that θ is not minimal or that $\mathcal{B}(\theta)$ is not controllable. Then there is exists a system of complexity (m, n') with $n' < n$ and with isometric transfer functions $(\hat{G}_0, \tilde{G}_0)$ such that

$$\int_{-\pi}^{\pi} \tilde{G}_0^*(e^{i\omega})\Sigma(e^{i\omega})\hat{G}_0(e^{i\omega})\frac{\alpha - e^{i\omega}}{1 - \alpha e^{i\omega}}d\omega = 0$$

holds for all $0 \neq \alpha \in (-\rho, \rho)$.

This is proved as follows. The error $e(\theta) = e(\mathcal{B}(\theta))$ depends only on the controllable part of $\mathcal{B}(\theta)$. Thus if θ is not minimal or if $\mathcal{B}(\theta)$ is not controllable, then there also exists an optimal system $\mathcal{B}_0$ of complexity (m, n') with $n' < n$. Let

$$\theta_0 = \left(\begin{array}{ccc} A_0 & B_0 & \tilde{B}_0 \\ C_0 & D & \tilde{D} \end{array} \right), \quad \theta_1 = \left(\begin{array}{cccc} \alpha & 0 & -\beta\gamma'T & 0 \\ \beta B_0\gamma & A_0 & \alpha B_0 T & \tilde{B}_0 \\ \beta D\gamma & C_0 & \alpha DT & \tilde{D} \end{array} \right)$$

where $\theta_0 \in \mathsf{R}^{(q+n-1)\times(q+n-1)}$ is an isometric representation of $\mathcal{B}_0$ with state dimension $n-1$ and with corresponding isometric transfer functions $\hat{G}_0(z^{-1}) = C_0(zI - A_0)^{-1}B_0 + D$ and $\tilde{G}_0(z^{-1}) = C_0(zI - A_0)^{-1}\tilde{B}_0 + \tilde{D}$. Further, $\theta_1 \in \mathsf{R}^{(q+n)\times(q+n)}$ is unitary for every $0 \neq \alpha \in (-1, 1)$, $\beta = \sqrt{1 - \alpha^2}$, $\gamma = (1, 0, \ldots, 0) \in \mathsf{R}^m$ and $T = \mathrm{diag}(1, 1/\alpha, \ldots, 1/\alpha)$. It follows from straightforward calculations that the system $\mathcal{B}(\theta_1)$ has isometric transfer functions $\hat{G}(z^{-1}) = \hat{G}_0(z^{-1})\,\mathrm{diag}((\alpha - z^{-1})/(1 - \alpha z^{-1}), 1, \ldots, 1)$ and $\tilde{G}(z^{-1}) = \tilde{G}_0(z^{-1})$. Therefore $P_0 = \hat{G}_0\hat{G}_0^* = \hat{G}\hat{G}^* = P$, so that the system $\mathcal{B}(\theta_1)$ is also optimal. By repeating the above reasoning for the other unit vectors γ the result follows from step 1.

Step 3. If $f(z)$ is a rational function that is analytic in an annulus containing the unit circle and if $\int_{-\pi}^{\pi} f(e^{i\omega})(\alpha - e^{i\omega})/(1 - \alpha e^{i\omega})d\omega = 0$ for all $0 \neq \alpha \in (-\rho, \rho)$, then $f(z)$ is analytic for all $|z| \leq 1$.

This is proved as follows. First we write $\int_{-\pi}^{\pi} f(e^{i\omega})(\alpha - e^{i\omega})/(1 - \alpha e^{i\omega})d\omega = \oint \phi(z)g(z)dz$ where $\oint$ denotes the integral along the unit circle, $\phi(z) = f(z)/z$ and $g(z) = (\alpha - z)/(1 - \alpha z)$. Let z_i, $i = 1, \ldots, k$, denote the poles of $\phi(z)$ within the unit circle, and let n_i denote the multiplicity of these poles. Then $\phi(z)$ has a Laurent series expansion around z_i of the form $\phi(z) = \sum_{j=-n_i}^{\infty} m_{i,j}(z - z_i)^j$. The series expansion of $g(z) = (\alpha - z)/(1 - \alpha z)$ is given by $g(z) = \sum_{j=0}^{\infty} k_{i,j}(\alpha)(z - z_i)^j$ where the coefficients $k_{i,j}(\alpha)$ are given by $k_{i,0} = (\alpha - z_i)/(1 - \alpha z_i)$ and $k_{i,j} = (\alpha^2 - 1)\alpha^{j-1}/(1 - \alpha z_i)^{j+1}$ for $j > 0$. Thus the residue of $\phi(z)g(z)$ at the point z_i is given by $\sum_{j=-n_i}^{-1} m_{i,j}k_{i,-1-j}(\alpha)$, and by the residue theorem we obtain $0 = \oint \phi(z)g(z)dz = \sum_{i=1}^{k} \sum_{j=-n_i}^{-1} m_{i,j}k_{i,-1-j}(\alpha)$. As this sum is a rational function of α which is zero on an interval of positive length it follows that this expression is zero for all $\alpha \in \mathsf{C}$. The linear independence of the functions $k_{i,j}(\alpha)$ implies that $m_{i,j}$ must be zero for all $j < 0$, so that $\phi(z)$ has no pole for $|z| \leq 1$. As $f(z) = z\phi(z)$ this proves the result.

Step 4. We now prove Lemma 11. Let Σ be a rational spectral density such that each of the entries σ_{ij} has at least $2n$ poles withinin the unit circle that are not poles of the other entries σ_{kl}, $(k, l) \neq (i, j)$. Suppose that there would exist an optimal system that is not controllable or that does not have full complexity. It then follows from Steps 2 and 3 that there exists a pair $(\hat{G}_0, \tilde{G}_0)$ of isometric transfer functions of state dimension (m, n') with $n' < n$ such that $\tilde{G}_0^* \Sigma \hat{G}_0$ is analytic

within the unit circle. As the (i,j)-th entry $(\tilde{G}_0^* \Sigma \hat{G}_0)_{ij}$ is a linear combination of the terms $\tilde{g}_{ki}^* \sigma_{kl} \hat{g}_{lj}$ and as $\tilde{g}_{ki}$ and $\hat{g}_{lj}$ have at most $n' < n$ zeros, at most $2n' < 2n$ poles of σ_{kl} can be cancelled in this product. Thus $\tilde{g}_{ki}^* \sigma_{kl} \hat{g}_{lj}$ has at least one pole within the unit circle that is not present in the other terms, so this is also a pole of $(\tilde{G}_0^* \Sigma \hat{G}_0)_{ij}$. This contradicts that this function is analytic within the unit circle. This proves that optimal systems are controllable and have full complexity (m, n).

This concludes the proof of part (i).

(ii) Let $\delta > 0$ be given and let the DGP belong to the generic set in (i) satisfying properties P1, P2 (for $\delta/3$) and P3. For given $\theta^* \in \Theta_\rho^*$ we define an open neighbourhood $\mathcal{U}(\theta^*) := \{\theta \in \Theta_\rho; \|\theta - \theta^*\| < \epsilon\}$, where $\epsilon > 0$ is chosen such that for all $\|\theta - \theta^*\| < 2\epsilon$ the system $\mathcal{B}(\theta)$ is controllable and has full complexity and in addition $d(\mathcal{B}(\theta), \mathcal{B}(\theta^*)) < \delta/3$. Let $\Theta' = \bigcup\{\mathcal{U}(\theta^*); \theta^* \in \Theta_\rho^*\}$ and $\Theta'' = \Theta_\rho \setminus \Theta'$. Then Θ'' is a closed subset of Θ_ρ and is thus compact. The continuity of e implies that the minimum of $e(\theta)$ over Θ'' is strictly larger than the minimum over Θ'. The uniform convergence in Lemma 4(iii) implies that, almost surely and for N sufficiently large, the infimum of e_N over Θ'' is also strictly larger than the infimum over the set Θ'. The continuity result in Proposition 2(iii) implies that e_N has a minimum over the closure of Θ', and hence also over Θ_ρ. As all systems $\mathcal{B}(\theta)$ with θ in the closure of Θ' are controllable and have full complexity, this proves P1 and P3, and P2 also follows from our construction of Θ'. $\square$

Proof of Lemma 8

(i) As before, by N we denote the sample size and by n the state dimension of a linear system. For $\tilde{G}(z^{-1})$ define $\tilde{G}_k^+(z^{-1}) = \tilde{D} + \sum_{j=1}^{k-1} CA^{j-1}\tilde{B}z^{-j}$ and $\tilde{G}_k^-(z^{-1}) = \sum_{j=k}^{\infty} CA^{j-1}\tilde{B}z^{-j}$. The isometry condition implies that $A'A + C'C = I$ and $AA' + \tilde{B}\tilde{B}' \le I$, so that $\sum_{j=0}^{\infty}(A')^j C'CA^j \le I$ and $\sum_{j=0}^{\infty} A^j \tilde{B}\tilde{B}'(A')^j \le I$. It then follows that the truncation error is given by $a(k) := \int (\tilde{G}_k^-)^*(\tilde{G}_k^-)d\omega = \tilde{B}'(A')^{k-1}\left(\sum_{j=0}^{\infty}(A')^j C'CA^j\right)A^{k-1}\tilde{B} \le \tilde{B}'(A')^{k-1}A^{k-1}\tilde{B}$. Evidently $a(k+1) \le a(k)$, and further $\sum_{k=1}^{\infty} a(k) \le nI$ because for every $\gamma \in \mathsf{R}^{q-m}$ with $\|\gamma\| = 1$ there holds $\sum_{k=1}^{\infty} \gamma' a(k)\gamma = \sum_{j=0}^{\infty} \text{trace}(A^j \tilde{B}\gamma\gamma'\tilde{B}'(A')^j) \le \text{trace}\sum_{j=0}^{\infty} A^j \tilde{B}\tilde{B}'(A')^j \le n$. So the truncation error is bounded by $a(k) \le (n/k)I$.

From (14) and (21) it follows that $\{e_N^0(\theta)\}^2 - \{e(\theta)\}^2 = \int \text{trace } \tilde{G}^* \Delta_N \tilde{G}d\omega$ where $\Delta_N = S_N - \Sigma$. Let $\gamma \in \mathsf{R}^{q-m}$ with $\|\gamma\| = 1$ be an arbitrary vector and let $H = \tilde{G}\gamma$, $H_+ = \tilde{G}_k^+\gamma$ and $H_- = \tilde{G}_k^-\gamma$. To show the uniform convergence of $e_N^0(\theta)$ to $e(\theta)$ it suffices to prove that $\int H^* \Delta_N H d\omega = \int H_+^* \Delta_N H_+ d\omega + 2\int H_+^* \Delta_N H_- d\omega + \int H_-^* \Delta_N H_- d\omega$ converges uniformly (in θ) to zero. Here $\Delta_N(z) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}(\hat{,}_j - ,_j)z^j$, where $,_j = \mathrm{E}\, w(t)w'(t-j)$ are the covariances of the observed process and $\hat{,}_j$ the corresponding sample covariances given by $\hat{,}_j = \frac{1}{N}\sum_{t=1}^{N} w(t)w'(t-j)\delta(t-j)$

with $\delta(t) = 1$ for $1 \le t \le N$ and $\delta(t) = 0$ elsewhere. Let $\Delta_{N,k} = \sum_{j=-k+1}^{k-1}(\hat{}_j - {}_{,j})z^j$, then it follows that $\int H_+^* \Delta_N H_+ d\omega = \int H_+^* \Delta_{N,k} H_+ d\omega$. Using the earlier obtained bounds for the truncation error $a(k)$ and that $\int H_+^* H_+ d\omega \le \int H^* H d\omega = 2\pi$, it follows that

$$\left| \int H^* \Delta_N H d\omega \right| \le 2\pi \|\Delta_{N,k}\|_\infty + 2\pi \|\Delta_N\|_\infty (2\sqrt{n/k} + n/k) \qquad (27)$$

This bound is independent of θ, and we have to show that it converges almost surely to zero when $N \to \infty$. For this purpose we let the truncation k depend on the sample size as $k = (\log N)^4$. As Σ is bounded on the unit circle it follows from Brillinger (1975, Theorem 7.7.2) that $\limsup_N \|\Delta_N\|/(\log N) < \infty$ almost surley, so that the second term in (27) converges to zero. Concerning the first term in (27), it follows from Hannan and Deistler (1988, Theorem 5.3.2) that $\limsup_N \{\max_{|j|<k} \|\hat{}_j - {}_{,j}\| \sqrt{N/\log\log N}\} < \infty$ almost surely, in which case also $\limsup_N \|\Delta_{N,k}\|_\infty \sqrt{N/\log\log N}/k < \infty$ and also the first term converges to zero. This concludes the proof of (i).

(ii) The criterion $e_N^0(\theta)$ of (21) is equal to $e(\theta)$ in (14) if the spectrum Σ is replaced by the periodogram S_N. As S_N is evidently bounded on the unit circle, the reasoning in the proof of Proposition 2(ii) shows that e_N^0 is continuous, and with the compactness in Proposition 2(i) the result follows. $\square$

Proof of Theorem 9

(i) The same line of reasoning applies as in the proof of Theorem 5. Indeed, the conditions for that proof are the compactness of the domain Θ_ρ, the continuity of the limit function $e(\theta)$, and the uniform convergence of e_N to e on Θ_ρ. According to Proposition 2(i) the set Θ is also compact, and Lemma 8(i) states the required uniform convergence.

(ii) That $\mathbf{B}_N^{0*}$ is non-empty is evident from Lemma 8(ii). It is assumed that the optimal system has full complexity, so that $\theta^* \in \Theta^*$ is minimal with $\lambda_{max}(A) = \lambda < 1$. It follows from (i) that almost surely and for N sufficiently large $\theta_N^0 \in \Theta_N^{0*}$ is also minimal and that $\theta_N^0 \in \Theta_\rho$ where $\rho := \frac{1}{2}(1 + \lambda) < 1$. The result now follows by the same arguments as in the proof of Theorem 6, that is, it follows from (i) and the continuity of $\mathcal{B}(\theta)$ as a function of θ in a neighbourhood of θ^* within Θ_ρ. $\square$

References

[1] Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.

[2] Beghelli, S., R.P Guidorzi and U. Soverini, The Frisch scheme in dynamic system identification. *Automatica* 26, 1990, pp. 171-176.

[3] Brillinger, D.R. (1975). *Time Series : Data Analysis and Theory*. Holt, Rinehart and Winston, New York.

[4] Caines, P.E. (1988). *Linear Stochastic Systems*. Wiley, New York.

[5] Davis, M.H.A. and R.B. Vinter (1985). *Stochastic Modelling and Control*. Chapman and Hall, London.

[6] Deistler, M. (1989). Symmetric modelling in system identification. In H. Nijmeijer and J.M. Schumacher (Eds.), *Three Decades of Mathematical System Theory*. Lecture notes in control and information sciences vol. 135, pp.128-147. Springer, Berlin.

[7] Deistler, M. and W. Scherrer (1992). Identification of linear systems from noisy data. In D. Brillinger (Ed.), *New Directions in Time Series Analysis*, part II. IMA vol. 46, pp. 21-42. Springer, Berlin.

[8] Hannan, E.J. (1970). *Multiple Time Series*. Wiley, New York.

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

[10] Heij, C. (1989). *Deterministic Identification of Dynamical Systems*. Lecture notes in control and information sciences vol. 127. Springer, Berlin.

[11] Heij, C. and W. Scherrer (1994). On the consistency of identification by dynamic factor models. *Proceedings 33rd CDC*, pp. 2880-2885. IEEE, Florida.

[12] Heij, C. and W. Scherrer (1995). Consistency of global total least squares in stochastic system identification. *Report 9522/A*, Econometric Institute, Erasmus University Rotterdam, The Netherlands. (Also *Proceedings 13th IFAC World Congress*, San Francisco, 1996.)

[13] Heij, C., W. Scherrer and M. Deistler (1995). System identification by dynamic factor models. *Report 9501/A*, Econometric Institute, Erasmus University Rotterdam, The Netherlands. Submitted to SIAM Journal on Control and Optimization.

[14] Ljung, L. (1987). *System Identification : Theory for the User*. Prentice-Hall, Englewood Cliffs, New Jersey.

[15] Lütkepohl, H. (1993). *Introduction to Multiple Time Series*. Springer, Berlin.

[16] Priestley, M.B. (1981). *Spectral Analysis and Time Series* (two vol.). Academic Press, London.

[17] Roorda, B. (1995a). *Global Total Least Squares*. Tinbergen Institute Research Series, vol. 88. Thesis Publishers, Amsterdam.

[18] Roorda, B. (1995b). Algorithms for global total least squares modelling of finite multivariable time series. *Automatica* 31, pp. 391-404.

[19] Roorda, B. and C. Heij (1995). Global total least squares modelling of multivariable time series. *IEEE Transactions on Automatic Control* 40, pp. 50-63.

[20] Willems, J.C. (1986, 1987). From time series to linear system (three parts). *Automatica* 22, 23, pp. 561-580, 675-694, 87-115.

[21] Willems, J.C. (1991). Paradigms and puzzles in the theory of dynamical systems. *IEEE Transactions on Automatic Control* 36, pp. 259-294.