System Identification by Dynamic Factor Models[*]

C. Heij [†] W. Scherrer [‡] M. Deistler [‡]

revised version, June 5, 1996

Abstract

This paper concerns the modelling of stochastic processes by means of dynamic factor models. In such models the observed process is decomposed into a structured part called the latent process, and a remainder that is called noise. The observed variables are treated in a symmetric way, so that no distinction between inputs and outputs is required. This motivates the condition that also the prior assumptions on the noise are symmetric in nature. One of the central questions in this paper is how uncertainty about the noise structure translates into non-uniqueness of the possible underlying latent processes. We investigate several possible noise specifications and analyse properties of the resulting class of observationally equivalent factor models. This concerns in particular the characterization of optimal models and properties of continuity and consistency.

Keywords: linear systems, stationary processes, identification, factor analysis, errors in variables, least squares, consistency.

[*] Support by the Austrian "Fonds zur Förderung der wissenschaftlichen Forschung" Projekt $P9176-PHY$, and by the EC-SCIENCE "System Identification" Project SC1*-CT92-0779 is gratefully acknowledged.

[†] Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, phone +31-10-4081269, fax +31-10-4527746.

[‡] Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Argentinierstrasse 8, A-1040 Wien, Austria, phone +43-1-58801-4405, fax +43-1-5054524.

1 Introduction

In this paper we are concerned with the identification of linear systems. The most commonly used models in system identification are ARMA and ARMAX models, we refer to [17], [4] and [13]. An ARMA model is symmetric and non-open, in the sense that all observed variables are treated in a symmetric way and that they are completely described by the model. On the other side, ARMAX models are non-symmetric and open, as a distiction is made between inputs and outputs and the noise is added to the outputs, and the inputs are not modelled.

We will consider linear factor models where the noise model is symmetric and where we have a deterministic, symmetric and open system model. In a sense these models combine the symmetry which is inherent in, for example, ARMA models, with the flexibility of models that leave certain process aspects unexplained, as for example in input-output models.

Of course, the classical ARMA and ARMAX models are appropriate in a great number of cases. For instance, if we are interested in predicting the outputs from the inputs then the ARMAX setting is appropriate. On the other hand, there are also situations where this approach can not be justified and may lead to prejudiced results.

- A prediction based error model is not appropriate, for example, if we are interested in the 'true' underlying system and there is noise on the inputs and the outputs.

- There may be uncertainty about the number of system equations or about the classification of the system variables into inputs and outputs. In this case we have to perform a more symmetric way of modelling, which in turn demands a symmetric noise model.

- In multivariate time series analysis one is confronted with the so-called curse of dimensionality. One method of reducing the dimension of the parameter space is dynamic factor analysis, which is an essential aspect of the approach described here.

Factor models have been used in statistics, psychometrics and econometrics for a long time, see [9], [1], [10]. The theory is most well-developed for the case of static models. Most applications are also reported within this framework, although there are also contributions on the identification of dynamic factor models, see [11], [8], [5]. Within the area of systems and control there is recently an increasing interest in symmetric modelling. We mention the introduction of the behavioural approach in systems theory in [24], [26], the attention for the Frisch problem, see [18], [23], [2], and low-noise modelling as proposed in [15]. Most contributions on factor models in this area deal either with the mathematical structure of dynamic models or with data modelling by means of static models. In an, in a certain sense, nonparametric framework results on the identification of dynamic factor models within the setting of stochastic errors in variables models have been presented in [6], [7]. Procedures for symmetric time series modelling within a deterministic behavioural framework have been proposed in [25], [14], and [21].

In this paper we try to integrate the above two frameworks, i.e., stochastic factor models and deterministic behavioural modelling. The model class consists of stochastic dynamic factor models where the latent process satisfies deterministic behavioural laws. This means that stochastic structure is added to the deterministic behavioural framework, which provides additional tools of analysis. On the other hand, our approach allows for an analysis of dynamic factor models in terms of finite dimensional systems, as opposed to the nonparametric results that were previously obtained.

We consider a situation which is idealized in so far as we commence from the population second moments of the data. In other words, we analyse the relation between the spectral density of the observed process and the corresponding factor models. Nevertheless, this is done from the point of view of requirements connected with the identification from observed data, and we will indicate how the results of this paper can be used for this purpose. A detailed analysis of procedures for the identification of dynamic factor models from observed time series falls beyond the scope of this paper and will be investigated elsewhere.

One of the issues studied in this paper is the non-uniqueness of the behaviour for given second order moments. This means that uncertainty about the precise noise structure leads to a corresponding non-uniqueness of the possible factor models that are compatible with the observed process. As is well known, in the main stream approach of modelling with exogenous inputs the population second moments of the observations determine, under very general conditions, the transfer function of the underlying system uniquely. This is due to the assumption that the noise is uncorrelated with the inputs. Uniqueness in general does not hold true in case all the variables may be corrupted by noise. This means that the set of observationally equivalent models, that is, the set of all models compatible with the population second moments, will in general not be a singleton. Of course, by imposing sufficiently strong conditions uniqueness can be achieved, but in many cases it may be hard to justify such assumptions. The question then becomes how the lack of knowledge about the error structure translates into non-uniqueness of the resulting model. This is a kind of uncertainty about the underlying system that can not be removed, even in an infinite sample.

We now give an outline of the topics treated in this paper. A dynamic factor model is of the form

$$w = \hat{w} + \tilde{w} \tag{1}$$

where w is the observed process, $\hat{w}$ is an (in general unobserved) latent process satisfying exact linear dynamic equations, and $\tilde{w}$ is the noise process. These restrictions can be expressed in terms of deterministic system behaviours as introduced in [24], [26]. The processes $(w, \hat{w}, \tilde{w})$ are assumed to be jointly stationary, and in this case the latent process has a singular spectrum. The noise process represents the error resulting from the approximation of the observations w by the latent process $\hat{w}$.

The central question considered in this paper is how to obtain the restrictions satisfied by the latent process from the observations. Without imposing further conditions, no solutions can be excluded from the knowledge of the observed process alone. This means that we have to impose additional assumptions on the noise structure in order to make meaningful statements about the underlying system. The main topics of this paper can be summarized as follows.

(i) The formulation of noise assumptions and an analysis of their effect on the class of observationally equivalent models. We consider in particular the assumptions of orthogonality (the latent process and the noise process are mutually uncorrelated), observability (the latent process can be expressed as a linear function of the observed process), and bounded noise (the noise process satisfies an a priori specified bound).

(ii) An analysis of the structural properties of identification procedures corresponding to different noise assumptions. This involves an analysis of the mapping relating an observed process to the class of observationally equivalent models. Continuity of this mapping is related to consistency in case of modelling from observed time series.

(iii) An analysis of the complexity and goodness of fit of factor models, with special attention for optimal models of restricted complexity.

This paper has the following structure. In Section 2 we define the dynamic factor model. For this purpose we review the behavioural approach in linear system theory. Factor models are characterized on the behavioural level and also in terms of spectral properties, and we define the complexity and goodness of fit of factor models. The general framework is illustrated by the special case of a white noise process and non-dynamic system equations, and it is shown that in this case our set-up coincides with the classical formulation of static factor models. Section 3 is concerned with optimal models, in the sense of minimizing the noise under restrictions on the compexity of the latent process. Section 4 investigates structural properties of the corresponding identification problem, with special attention for continuity and consistency. Section 5 contains concluding remarks. Some technical proofs are collected in the appendix.

2 Dynamic Factor Models

2.1 Linear Systems

For the formulation of dynamic factor models it is convenient to use the behavioural approach as developed by Willems in [24], [26]. Since this approach may be not well-known to the reader, we discuss in this section those aspects that are relevant for our purposes. Readers with an interest for further details and proofs are referred to [24], [26].

In this subsection $\hat{w} : \mathbf{Z} \to \mathbf{R}^q$ denotes a trajectory rather than a process, that is, it is a q-variate time series observed in discrete time. The behaviour of a deterministic system is defined as the set of all trajectories $\hat{w}$ that may arise within the restrictions imposed by the system. So a behaviour is a subset $\mathcal{B}$ of $(\mathbf{R}^q)^{\mathbf{Z}}$. Of special interest are behaviours that are linear, time invariant, and complete. This means that $\mathcal{B} \subset (\mathbf{R}^q)^{\mathbf{Z}}$ is a linear subspace that is invariant under the shift operator σ, defined by $(\sigma\,\hat{w})(t) := \hat{w}(t+1)$, 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 $\mathbf{R}^q$) to $\hat{w}_0 \in (\mathbf{R}^q)^{\mathbf{Z}}$ there holds that also $\hat{w}_0 \in \mathcal{B}$. These conditions imply that the behaviour corresponds to a linear, time invariant, finite dimensional system. In the sequel we will simply use the term linear system to refer to a linear, time invariant, complete behaviour $\mathcal{B} \subset (\mathbf{R}^q)^{\mathbf{Z}}$.

Linear systems can be represented in several ways. Here we discuss representations in terms of polynomial equations, state space models with driving variables, and corresponding transfer functions.

Every linear system can be represented in polynomial form, as the solution set of the polynomial equations

$$R(\sigma, \sigma^{-1})\,\hat{w} = 0 \tag{2}$$

Here R is a polynomial matrix in the forward and backward shifts. The representation of a given system by a polynomial matrix is highly non-unique. Without loss of generality we could have restricted ourselves to polynomials in either σ or in σ^{-1} alone, but (2) is in accordance with [24], [26]. The set of behavioural laws of a linear system $\mathcal{B}$ is defined as the set of all polynomial equations satisfied by the system, that is, it is the module of $1 \times q$ polynomials $\mathcal{L} = \{r; r(\sigma, \sigma^{-1})\,\hat{w} = 0 \text{ for all } \hat{w} \in \mathcal{B}\}$. Every polynomial representation of a given system has the same (polynomial) rank p, which is equal to the dimension of the module $\mathcal{L}$. Full row rank representations are unique up to left multiplication by a unimodular matrix, i.e., a polynomial matrix which has a polynomial inverse. These representations can also be interpreted as input-output systems in polynomial form, where p is the number of outputs and $m := q - p$ is the number of inputs. We denote by n the minimal number of initial conditions required to express future outputs in terms of future inputs, which is equal to the sum of the Kronecker observability indices of the system.

An alternative representation is in terms of state models with driving variables. Every linear system can be represented as

$$\sigma x = Ax + Bv, \quad \hat{w} = Cx + Dv \tag{3}$$

Here v is an auxiliary vector of unrestricted driving variables and x is a vector of state variables. In contrast with the usual input-state-output model, here all the external variables are described as outputs of a system driven by forces which need not have any external meaning. For a given system this kind of representation is highly non-unique. Minimal representations have n states and m driving variables, and the class of all minimal representations is described by the feedback group $(S(A+BF)S^{-1}, SBR, (C+DF)S^{-1}, DR)$.

Until now no assumptions were made concerning the controllability of systems. For example, if A is a $q \times q$ invertible matrix then the set $\{\hat{w} : \mathbf{Z} \to \mathbf{R}^q; \hat{w}(t+1) = A\,\hat{w}(t), t \in \mathbf{Z}\}$ defines a linear system with autonomous evolution which is clearly not controllable. A system $\mathcal{B}$ is called controllable if every future in $\mathcal{B}$ is attainable from every past in $\mathcal{B}$, that is, if for every $\hat{w}_1, \hat{w}_2 \in \mathcal{B}$ there exist $\hat{w} \in \mathcal{B}$ and $h \geq 0$ such that $\hat{w}(t) = \hat{w}_1(t)$ for $t < 0$ and $\hat{w}(t) = \hat{w}_2(t)$ for $t \geq h$. In terms of the kernel representations (2) this means that $R(z, z^{-1})$ has constant rank over $z \in \mathbf{C} \setminus \{0\}$. In this case the system can also be represented as the image of a polynomial operator, that is, $\hat{w} \in \mathcal{B}$ is represented as

$$\hat{w} = M(\sigma, \sigma^{-1})f \tag{4}$$

where f has the interpretation of the underlying generating factors. There is a close connection between the notion of controllability as defined before and the usual notion in terms of state space models, because minimal state models (3) of controllable systems $\mathcal{B}$ are characterized by the property that (A, B) is a controllable pair and (A, C) an observable pair. In this case we can obtain isometric state models, see [21], that is, representations with the property that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}' \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I_n & O \\ O & I_m \end{pmatrix} \tag{5}$$

where I_d denotes the d-dimensional identity matrix and Q' denotes the transposed of a matrix Q. If (A, B, C, D) is a minimal isometric state representation of a controllable system, then all such representations are given by $(UAU'$, UBV, CU', $DV)$ with U and V orthogonal matrices.

The model (4) gives a finite impulse response representation of controllable systems. This gives a clear description how to generate all time series belonging to a given system. Alternative descriptions are in terms of transfer functions. For controllable systems we can always choose A to be asymptotically stable, and in this case the square summable time series in the system can be generated as $\hat{w} = G(\sigma^{-1})v$, where v is square summable and G is the causal transfer function defined by $G(z) = D + \sum_{k=1}^{\infty} CA^{k-1}Bz^k$. The rank of the transfer function G is m, and its McMillan degree is n. For an isometric state model this transfer function becomes an isometry, sometimes also called an all-pass transfer function. The driving variables needed to generate a given square summable time series are then obtained by $v = G^*(\sigma^{-1})\,\hat{w}$, where G^* is the adjoint defined by $G^*(\sigma^{-1}) := G'(\sigma)$.

In our analysis we will often make use of isometric representations of linear systems. A state space method for obtaining these models is described in [21]. They can also be obtained from polynomial representations, as follows. Let $\mathcal{B}$ be a controllable linear system with kernel representation $\mathcal{B} = \ker(R) = \{\hat{w}; R(\sigma,\sigma^{-1})\,\hat{w} = 0\}$ and image representation $\mathcal{B} = \mathrm{im}(M) = \{\hat{w}; \hat{w} = M(\sigma,\sigma^{-1})f\}$. If m is the number of inputs of the system, then R can be chosen with $q-m$ rows and M with m columns. Controllability implies that $R(z,z^{-1})$ has constant rank over $\mathbf{C}\setminus\{0\}$, and M can also be chosen of constant rank. In this case the projections $P = M(M^*M)^{-1}M^*$ and $Q = R^*(RR^*)^{-1}R$ are well-defined rational functions with constant rank over the domain $\mathbf{C}\setminus\{0\}$. So there exist causal, miniphase spectral factorizations $P = \hat{G}\,\hat{G}^*$ and $Q = \tilde{G}\,\tilde{G}^*$, see [22, theorem I.10.1]. These spectral factors are isometric, that is, $\hat{G}^*\,\hat{G} = I_m$ and $\tilde{G}^*\,\tilde{G} = I_{q-m}$. Then the spectral factor $\hat{G}$ is an isometric transfer function for $\mathcal{B}$, and all square summable time series in $\mathcal{B}$ are obtained as the image of $\hat{G}$. Therefore we call this an isometric image representation. Further, all square summable time series in $\mathcal{B}$ are annihilated by $\tilde{G}^*$ and therefore we call $\tilde{G}$ an isometric kernel representation. As R and M describe the same system, it follows that $RM = 0$ so that $\tilde{G}^*\,\hat{G} = 0$. This shows that the $q \times q$ rational matrix $[\hat{G},\tilde{G}]$ is inner, that is, it is stable and unitary. Conversely, every rational inner matrix $[\hat{G},\tilde{G}]$ describes a linear system with isometric image representation $\hat{G}$ and isometric kernel representation $\tilde{G}$.

2.2 Factor Models and Spectra

Let $(\Omega, \mathcal{A}, \mathrm{P})$ denote an underlying probability space and let $\mathbf{L}_2$ be the corresponding Hilbert space of square integrable real-valued random variables. We assume that the observed process w consists of q-dimensional random vectors, so that $w \in (\mathbf{L}_2^q)^{\mathbf{Z}}$. A dynamic factor model is a process decomposition of the form $w = \hat{w} + \tilde{w}$, where $\tilde{w} \in (\mathbf{L}_2^q)^{\mathbf{Z}}$ is the noise process and $\hat{w} \in (\mathbf{L}_2^q)^{\mathbf{Z}}$ is the latent process that is essentially restricted to a linear system. The behaviour $\mathcal{B}$ of $\hat{w}$ is defined as the smallest linear, time invariant, complete system which contains almost all process realizations, that is, $P\{\hat{w}(\omega) \in \mathcal{B}\} = 1$. The following result states that this definition makes sense.

Proposition 1 *For every stochastic process the behaviour is well-defined.*

Proof. We call a behaviour $\mathcal{B}$ compatible with a process $\hat{w}$ if $\mathcal{B}$ contains almost all process realizations. Of course $(\mathbf{R}^q)^{\mathbf{Z}}$ is always compatible, and countable intersections of compatible behaviours are compatible.

Now let $\mathcal{B}$ be a compatible behaviour. If it contains a strictly smaller compatible behaviour $\mathcal{B}' \subset \mathcal{B}$, $\mathcal{B}' \neq \mathcal{B}$, then we proceed with $\mathcal{B}'$. This system has either less inputs than $\mathcal{B}$, or it has equal number of inputs and less states. Continuing in this way, we end up after a finite number of steps with a compatible behaviour $\mathcal{B}^*$ that contains no strictly smaller compatible behaviour. This implies that for every compatible $\mathcal{B}$ there holds $\mathcal{B} \cap \mathcal{B}^* = \mathcal{B}^*$, and thus $\mathcal{B}^* \subseteq \mathcal{B}$. This proves that $\mathcal{B}^*$ is the smallest compatible behaviour. □

We call a behaviour nontrivial if $\mathcal{B} \neq (\mathbf{R}^q)^{\mathbf{Z}}$. Dynamic factor models are defined as follows.

Definition 1 *A dynamic factor model of a process w is a decomposition $w = \hat{w} + \tilde{w}$ where the latent process $\hat{w}$ has nontrivial behaviour $\mathcal{B}$, which is called the behaviour of the factor model.*

In this paper we will be mainly concerned with the behaviour of factor models, as in many cases this is the main point of interest in system identification. In order to simplify our analysis of dynamic factor models we make some additional assumptions on the processes. Some of these assumptions could be relaxed, but they are imposed to prevent technical complications that could obscure the underlying modelling ideas. To formulate the assumptions we use the following terminology. Let S_t denote the subspace of $\mathbf{L}_2$ spanned by the zero mean random variables $\{w_i(t); i = 1, \cdots, q\}$. Let the Hilbert spaces $\mathbf{H}(w)$ and $\mathbf{H}_t(w)$ be generated by respectively $\{S_t; t \in \mathbf{Z}\}$ and $\{S_s; s \leq t\}$, so that $\mathbf{H}(w)$ is generated by the process and $\mathbf{H}_t(w)$ by the past of this process. The process is said to have full rank if the space $\mathbf{H}_t(w) \cap \{\mathbf{H}_{t-1}(w)\}^{\perp}$ has dimension q, that is, if no nontrivial linear combination of the variables $w(t)$ can be predicted without error from the past. It is called purely nondeterministic if $\bigcap_{-\infty}^{\infty} \mathbf{H}_t(w) = \{0\}$, that is, if the prediction of $w(t+h)$ from $\mathbf{H}_t(w)$ converges to zero for $h \to \infty$. As is well known, every purely nondeterministic process can be written as

$$w = T(\sigma^{-1})\varepsilon \tag{6}$$

that is, $w(t) = \sum_{k=0}^{\infty} T_k \varepsilon(t-k)$ where ε is a white noise process with $\mathrm{E}\{\varepsilon(t)\varepsilon'(t)\} = I_q$ and $\varepsilon(t) \in \mathbf{H}_t(w)$ and where $\sum_{k=0}^{\infty} \|T_k\|_2^2 < \infty$. This is called a Wold representation of the process. If $\sum_{k=0}^{\infty} \|T_k\|_2 < \infty$ then this representation is called absolutely summable. In this paper we will always make the following assumptions.

Assumptions

- **A1** The processes w, $\hat{w}$ and $\tilde{w}$ are jointly weakly stationary, with zero mean and finite second order moments.

- **A2** The observed process w is purely nondeterministic and has full rank.

- **A3** The latent process $\hat{w}$ and the noise process $\tilde{w}$ are purely nondeterministic.
- **A4** The Wold representations of w, $\hat{w}$ and $\tilde{w}$ are absolutely summable.

The assumption A1 is imposed for convenience, as this means that the usual tools of time series analysis and linear systems theory become relevant. The full rank assumption in A2 implies that the behaviour of the observed process is unrestricted, so that it can not be modelled by a factor model without noise. Concerning assumption A3, note that a latent process with nontrivial behaviour can not be of full rank. We assume that it is purely nondeterministic, and that the same holds true for the noise. This seems a reasonable requirement in view of assumption A2. Finally, assumption A4 is imposed for technical reasons. It implies that the spectral densities of the processes are continuous functions on the unit circle.

Stated in terms of behaviours, assumption A3 for the latent process means the following.

Proposition 2 *The behaviour of a purely nondeterministic process is controllable.*

Proof. Let $\hat{w}$ be a purely nondeterministic process. Further let $\mathcal{B}$ be a non-controllable system with full row rank polynomial representation R, with the property that $R(\sigma, \sigma^{-1})\,\hat{w} = 0$ almost surely. Let $R = UDV$ be the Smith form, with U and V unimodular matrices and with $D = (\Delta, 0)$ where Δ is a diagonal matrix with one-dimensional polynomials unequal to zero on the diagonal.

Define $w^* = V\,\hat{w}$ and let $w^* = (w_1^*, w_2^*)$ be a partitioning corresponding to that of $D = (\Delta, 0)$. Then there holds $\Delta w_1^* = 0$ almost surely. So this process evolves according to an autonomous difference equation and can be predicted without error, that is, w_1^* belongs to $\mathbf{H}_t(\hat{w})$ for all t, the space spanned by the past of $\hat{w}$. As $\hat{w}$ is purely nondeterministic this means that $w_1^* = 0$. This shows that also $R^*(\sigma, \sigma^{-1})\,\hat{w} = 0$ almost surely, where $R^* = (I, 0)V$. As $R^*(z, z^{-1})$ has constant rank it follows that this defines a controllable system, and of course it defines a system that is strictly smaller than $\mathcal{B}$. So the behaviour of $\hat{w}$ is also controllable. □

We mention that the converse of this result does not hold true, that is, a latent process with controllable behaviour need not be purely nondeterministic. In terms of the representations of controllable systems discussed in Section 2.1, the above result means that a factor model can be described as follows.

$$w = M(\sigma, \sigma^{-1}) f + \tilde{w} \tag{7}$$

$$w = Cx + Dv + \tilde{w}, \quad \sigma x = Ax + Bv \tag{8}$$

Here M is a polynomial matrix and (A, B, C, D) are real-valued matrices. The first representation is a generalization of the static model of classical factor analysis and explains the observed variables in terms of a number of unobserved underlying factors. The second representation gives a more explicit description of the dynamical evolution of the latent process $\hat{w} = Cx + Dv$ in terms of unrestricted factors v and additional factors x that exhibit the memory structure.

Factor models can also be described by means of spectra. In terms of the Wold representation (6), where ε is white noise with unit covariance and where T is an (in general nonrational) causal transfer function with causal inverse, the spectrum of w is given by $\Sigma = TT^*$. The spectra of $\hat{w}$ and $\tilde{w}$ are denoted respectively by $\hat{\Sigma}$ and $\tilde{\Sigma}$, and the cross spectrum between $\hat{w}$ and $\tilde{w}$ is denoted by Σ_c. Under Assumptions A1-A4, all these spectra are bounded functions on the unit circle. A factor model corresponds to a decomposition

$$\Sigma = \hat{\Sigma} + \tilde{\Sigma} + \Sigma_c + {\Sigma_c}' \tag{9}$$

By assumption, the behaviour of the latent process is nontrivial so that $\hat{\Sigma}$ is singular. The rank of this spectrum corresponds to the number of unrestricted factor components. This is made precise in the following result. Here we denote by $\ker(\hat{\Sigma})$ the set of $1 \times q$ polynomials $r(s, s^{-1})$ for which $r(z, z^{-1})\,\hat{\Sigma}(z) = 0$ on the unit circle. The polynomial rank of $\hat{\Sigma}$ is defined as $q - p$, where p is the dimension of the module $\ker(\hat{\Sigma})$. Further, by $\mathrm{im}(\hat{\Sigma})$ we denote the smallest linear system that contains all time series of the form $\hat{\Sigma}(\sigma)v$, where v is a $q \times 1$ time series with finite support.

Theorem 3

(i) A latent process $\hat{w}$ with spectrum $\hat{\Sigma}$ has behaviour $\mathcal{B} = \mathrm{im}(\hat{\Sigma})$, and the behavioural laws are given by $\mathcal{L} = \ker(\hat{\Sigma})$.

(ii) The number of inputs of the behaviour is equal to the polynomial rank of $\hat{\Sigma}$.

(iii) A latent process has behaviour $\mathcal{B}$ if and only if it can be generated as $\hat{w} = \hat{G}\,v$, where $\hat{G}$ is an isometric image representation of $\mathcal{B}$ and v is a weakly stationary process with zero mean and finite second order moments that has trivial behaviour.

Proof. (i) Let $\mathcal{B}$ be the behaviour of $\hat{w}$ and $\mathcal{L}$ the corresponding set of laws. Then a $1 \times q$ polynomial belongs to $\mathcal{L}$ if and only if $r\,\hat{w} = 0$ holds almost surely, and this is equivalent to the condition $r\,\hat{\Sigma} = 0$, that is, $\mathcal{L} = \ker(\hat{\Sigma})$.

Now let $\mathcal{B}^* = \mathrm{im}(\hat{\Sigma})$ be the smallest linear system that contains all time series of the form $\hat{\Sigma}(\sigma)v$, where v is a $q \times 1$ time series with finite support. Let $\mathcal{L}^*$ denote the set of laws of the system $\mathcal{B}^*$. The system $\mathcal{B}^*$ consists of pointwise limits of time series $\hat{\Sigma}(\sigma)v_n$, $n = 1, 2, \ldots$ where v_n are time series with finite support. If $r \in \mathcal{L}$ then $r\,\hat{\Sigma} = 0$ implies $r(\sigma)\,\hat{\Sigma}(\sigma)v_n = 0$, and the same holds true for the pointwise limit of $\hat{\Sigma}(\sigma)v_n$. This shows that $\mathcal{L} \subseteq \mathcal{L}^*$. Now let r be a $1 \times q$ polynomial with $r\,\hat{\Sigma} \neq 0$ and let $w \in \mathcal{B}^*$ be defined by $w = \hat{\Sigma}(\sigma)v$ where v has Z-transform r'. As $r\,\hat{\Sigma}\,r' \neq 0$ it follows that $r(\sigma)\,\hat{\Sigma}(\sigma)v \neq 0$, so that r does not belong to $\mathcal{L}^*$. This implies that $\mathcal{L}^* \subseteq \mathcal{L}$, so that $\mathcal{L}^* = \mathcal{L}$. As $\mathcal{B}$ and $\mathcal{B}^*$ satisfy the same relations it follows that $\mathcal{B} = \mathcal{B}^*$.

(ii) The number of inputs of $\mathcal{B}$ is given by $m = q - p$, where p is the dimension of the module $\mathcal{L} = \ker(\hat{\Sigma})$. This was also defined as the polynomial rank of $\hat{\Sigma}$.

(iii) First assume that $\hat{w}$ has behaviour $\mathcal{B}$ with m inputs. Let R be a $(q - m) \times q$ polynomial matrix with full rank so that $\mathcal{B} = \ker(R)$, and let $\hat{G}$ be an isometric image representation of $\mathcal{B}$ as defined in Section 3.1, so that $R\,\hat{G} = 0$. As $\hat{G}$ is rational it can be written as $p^{-1}Q$, with p a scalar polynomial and Q a $q \times m$ matrix polynomial with full column rank. As $\hat{G}$ is stable, so that it has no poles on the unit circle, it follows that $\hat{v} = \hat{G}^*\,\hat{w}$ is a well-defined stationary process with zero mean and finite second order moments. As $\hat{G}\,\hat{G}^*$ is the projection onto $\mathcal{B}$ and realizations of the factor process belong almost surely to $\mathcal{B}$, it follows that $\hat{G}\,\hat{v} = \hat{G}\,\hat{G}^*\,\hat{w} = \hat{w}$. It remains to show that $\hat{v}$ has trivial behaviour $(\mathbf{R}^m)^{\mathbf{Z}}$. Suppose that this was not the case, then there is a $1 \times m$ polynomial $r \neq 0$ such that $r\,\hat{v} = 0$. As Q has rank m there exists a $1 \times q$ polynomial π so that $\pi Q = r$, and $\pi\,\hat{w} = p^{-1}\pi Q\,\hat{v} = 0$ so that π is a law of the process $\hat{w}$. It then follows that $(R', \pi')'\,\hat{G} = (0, r')'$ where $r \neq 0$. This implies that $(R', \pi')'$ is a polynomial matrix of rank $q - m + 1$ with the property that $(R', \pi')'\,\hat{w} = 0$. This means that the behaviour of $\hat{w}$ has less than m inputs, but this contradicts (ii).

Second, suppose that $\hat{w} = \hat{G}\,\hat{v}$. As $\hat{v}$ has trivial behaviour it follows that r is a behavioural law of $\hat{w}$ if and only if $r\,\hat{G} = 0$, or equivalently $r\,\hat{G}\,\hat{G}^* = rP = 0$ with P the projection operator onto $\mathcal{B}$. This shows that the behaviour of $\hat{w}$ is given by $\mathcal{B}$. □

Concerning (ii), note that the polynomial rank of $\hat{\Sigma}$ is $q - p$, where p is the number of independent polynomial relations satisfied by the latent process $\hat{w}$. In general, the polynomial rank may be larger than the dimension of the innovation space $\mathbf{H}_t(\hat{w}) \cap \{\mathbf{H}_{t-1}(\hat{w})\}^{\perp}$. This dimension is the usual definition of the rank of the process $\hat{w}$, and this is equal to the maximum of $\text{rank}(\hat{\Sigma}(z))$ on the unit circle. This implies that for all $|z| = 1$ the rank of $\hat{\Sigma}(z)$ is smaller than or equal to the polynomial rank of $\hat{\Sigma}$, and if $\hat{w}$ satisfies additional linear relations that are not polynomial then the rank of $\hat{\Sigma}(z)$ is strictly smaller than the polynomial rank of $\hat{\Sigma}$. As nonpolynomial relations correspond to infinite dimensional systems they fall outside the behavioural setting discussed in Section 2.1.

2.3 Factor Schemes

The basic question considered in this paper concerns the relationship between the spectrum of the observed process and the class of observationally equivalent factor models. Under Assumption A2 there exists for every linear system $\mathcal{B}$ a factor model with behaviour $\mathcal{B}$, because we can simply define the noise as $\tilde{w} = w - \hat{w}$ for every latent process $\hat{w}$. In the words of Kalman [15], within this setting we can obtain no models without prejudice. So we have to impose additional restrictions on the noise process in order to make meaningful statements about the underlying system. These restrictions should be motivated in each practical situation. Here we consider the following possible specifications, which we call factor schemes.

- The factor model is called orthogonal if the latent process and the noise process are mutually uncorrelated, that is, if $\mathrm{E}\{\hat{w}(t)\,\tilde{w}(s)'\} = 0$ for all t, s. Stated otherwise, there holds $\mathbf{H}(\hat{w}) \perp \mathbf{H}(\tilde{w})$ and $\Sigma_c = 0$.
- The factor model is called observable if $\hat{w}$ is a linear function of w, that is, if $\mathbf{H}(\hat{w}) \subseteq \mathbf{H}(w)$. Stated otherwise, there holds $\hat{\Sigma} = F\,\Sigma\,F'$, $\tilde{\Sigma} = (I-F)\,\Sigma(I-F)'$ and $\Sigma_c = F\,\Sigma(I-F)'$ for some, possibly noncausal, transfer function F.
- The factor model is said to have bounded noise if it satisfies an a priori specified bound in terms of the noise spectrum $\tilde{\Sigma}$.

The quality of factor models is expressed in terms of the complexity and the goodness of fit of the model.

Definition 2 *The* complexity *of a dynamic factor model is defined as the pair (m, n), where m is the number of driving variables and n the number of states of the behaviour of the factor model.*

The complexity measures the dimension of the latent process, in the sense that the set of possible realizations $\{\hat{w}(\omega); \omega \in \Omega\}$ on a time interval of length $L \geq n$ is (almost surely) contained in an $(mL + n)$-dimensional subspace of $\mathbf{R}^{qL}$. In parametric terms, the complexity can also be expressed as follows.

Proposition 4

(i) In terms of a kernel representation $R(\sigma, \sigma^{-1})\,\hat{w} = 0$, the complexity is given by $m = q - \mathrm{rank}(R)$ and $n = \sum_{k=1}^{q-m} \nu_k$, where $\{\nu_1, \cdots, \nu_{q-m}\}$ are the Kronecker observability indices.

(ii) In terms of an isometric image representation $\hat{G}$ of the factor behaviour, the complexity is given by the rank m and McMillan degree n of $\hat{G}$.

Proof. (i) This follows from Theorem 6 in [24].

(ii) This follows from Theorem 4.9 and Lemma 4.10 in chapter 4 of [14]. □

In the sequel we will sometimes consider another measure of complexity in case the factor model is observable and the spectrum Σ is rational, that is,

$w = T(\sigma^{-1})\varepsilon$ in (6) where T is now a rational transfer function. Then a special class of latent processes is obtained by prefiltering the noise, that is, $\hat{w} = T(\sigma^{-1})F(\sigma^{-1})\varepsilon$ where F is a rational, rank deficient transfer function. We define the effective noise space by $\mathcal{N}$ =im(F), that is, the behaviour of the filtered noise process $F(\sigma^{-1})\varepsilon$. In this case the behaviour of the latent process is given by $\mathcal{B}$ =im(TF). As TF is rational and rank deficient, it follows that $\mathcal{B}$ is a nontrivial linear system. An alternative characterization of complexity is the pair (m', n'), the number of inputs and states of the effective noise behaviour $\mathcal{N}$. This measures the complexity of the noise process underlying the latent process.

Definition 3 *Let be given a process with rational Wold representation $w = T(\sigma^{-1})\varepsilon$ and a latent process $\hat{w} = (TF)(\sigma^{-1})\varepsilon$. Then the* noise complexity *of the corresponding factor model is defined as (m', n'), the number of inputs and states of the effective noise space $\mathcal{N} = \mathrm{im}(F)$.*

The two foregoing notions of complexity are not equivalent. If $\mathcal{N}$ is the effective noise space of a factor model with behaviour $\mathcal{B}$ and if (m, n) and (m', n') are the complexities of $\mathcal{B}$ and $\mathcal{N}$ respectively, then $m = m'$ but in general $n \neq n'$.

The goodness of fit of factor models is measured in terms of the second moments of the noise process $\tilde{w}$. As is well known, the choice of norms may have an essential effect on the obtained models. Here we will restrict the attention to the mean squares norm and the uniform norm. In the following we use the notation $\tilde{\Sigma}^{1/2}$ for a spectral factor of the noise spectrum $\tilde{\Sigma}$ so that $\tilde{\Sigma} = \tilde{\Sigma}^{1/2}(\tilde{\Sigma}^{1/2})^*$. We define the norm of a $1 \times q$ polynomial $r(\sigma, \sigma^{-1}) = \sum r_k \sigma^k$ by $\|r\|_2^2 := \sum \|r_k'\|^2$ where $\| \cdot \|$ denotes the Euclidean norm on $\mathbf{R}^q$. Further we define the following norms for spectral factors, where $\lambda_{\max}(Q)$ denotes the spectral radius, that is, the maximum of the absolute values of the eigenvalues of a matrix Q.

$$\| \tilde{\Sigma}^{1/2} \|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{trace}\{\tilde{\Sigma}(e^{-i\lambda})\} d\lambda \tag{10}$$

$$\| \tilde{\Sigma}^{1/2} \|_\infty^2 = \sup_{\lambda \in [-\pi, \pi]} \lambda_{\max}\{\tilde{\Sigma}(e^{-i\lambda})\} \tag{11}$$

Definition 4 *For a factor model with noise process $\tilde{w}$ with spectrum $\tilde{\Sigma}$, the mean squares and uniform fit are respectively defined by*

$$\| \tilde{w} \|_2 := [\mathrm{E}\{\tilde{w}(t)' \, \tilde{w}(t)\}]^{1/2} = \| \tilde{\Sigma}^{1/2} \|_2 \tag{12}$$

$$\| \tilde{w} \|_\infty := \sup\{[\mathrm{E}\{(r(\sigma, \sigma^{-1}) \, \tilde{w})(t)\}^2]^{1/2}; \|r\|_2 = 1\} = \| \tilde{\Sigma}^{1/2} \|_\infty \tag{13}$$

Because of Assumption A3, the noise process is purely nondeterministic so that the coefficients of $\tilde{\Sigma}^{1/2}$ are square summable so that $\| \cdot \|_2$ is well-defined, and because of Assumption A4 the spectrum is bounded on the unit circle so that $\| \cdot \|_\infty$ is also well-defined. The mean squares and uniform norms are monotonic, as they become larger if the spectrum becomes larger in the sense of positive semidefinite matrix functions on the unit circle. Sometimes, when results hold true for both norms, we make no distinction in notation and write $\| \tilde{w} \|$ and $\| \tilde{\Sigma}^{1/2} \|$.

2.4 Illustrations

2.4.1 Static Factor Models

As a simple illustration we show that the framework as introduced before is an extension to the dynamic case of the well-known class of static factor models that have been analysed, among others, in [18] and [23]. In later sections we will use the static case for further illustration.

Suppose that the observations are uncorrelated over time, so that w is a white noise process. In this section we restrict the attention to factor models $w = \hat{w} + \tilde{w}$ where $\hat{w}$ and $\tilde{w}$ are also white noise processes. We further impose the condition that the behaviour of the factor model is static in the sense that the state dimension is $n = 0$. The corresponding linear systems are described by linear nondynamic equations of the form $R\hat{w} = 0$, where R is a full row rank $p \times q$ real matrix. Let M be a $q \times (q-p)$ matrix with $\mathrm{im}(M) = \ker(R)$, then the factor model can be written as

$$w(t) = Mf(t) + \tilde{w}(t).$$

This corresponds to the classical static factor model with factors f. If the covariance matrix of f has full rank, then the complexity of this factor model is $(m, 0)$, where $m = q - p$ is the number of factors.

In the literature several possible factor schemes have been proposed. For example, in the principal component analysis of multivariate statistics the aim is to keep the noise process $\tilde{w}$ as small as possible, under a restriction on the number of independent factors m. In the so-called Frisch-scheme the aim is to minimize the complexity of the model under the restrictions that the processes $\hat{w}$ and $\tilde{w}$ are orthogonal and that in addition the q components of the noise process $\tilde{w}$ are mutually orthogonal.

Our approach resembles principal component analysis, as in the next section we will consider minimization of the noise under a restriction of the complexity of the behaviour of the latent process.

2.4.2 Dynamic System Example

Here we give a simple example of a dynamic factor model. Suppose that the data generating process consists of a single input, single output system where both the input u and the output y are observed under additive noise. That is, we assume that the data $w = (u, y)$ are generated as $w = \hat{w} + \tilde{w}$, with $\tilde{w}$ the noise and $(\hat{u}, \hat{y})$ the latent process with $\hat{y} = g\,\hat{u}$ where g denotes the underlying rational transfer function. For simplicity we assume that the latent input $\hat{u}$ is white noise and that the noise process $\tilde{w}$ is also white noise, all uncorrelated and with unit variance. In this case the spectrum of the data generating process is given by

$$\Sigma = \begin{pmatrix} 2 & g^* \\ g & gg^* + 1 \end{pmatrix}$$

An obvious factor model for this process is the above decomposition in the latent process $\hat{w}$ and the white noise process $\tilde{w}$. If $g(\sigma, \sigma^{-1}) = r_1(\sigma, \sigma^{-1})/r_2(\sigma, \sigma^{-1})$ then this factor model has a behaviour described by the equation $R(\sigma, \sigma^{-1})\,\hat{w} = 0$ where $R = (-r_1, r_2)$. The complexity is $(m, n) = (1, d)$, where d is the maximum of the degrees of the polynomials r_1 and r_2. The mean squares fit is $\| \tilde{w} \|_2 = \sqrt{2}$ and uniform fit $\| \tilde{w} \|_\infty = 1$. Because of our assumptions, this factor model is orthogonal but not observable.

Of course, the real question is whether we can identify the underlying transfer function g from the spectrum Σ. This will be investigated in Section 3.3.2.

3 Pareto Optimal Models

The quality of a factor model for an observed process w is measured by its complexity and goodness of fit. In general, the fit can become better if the model is allowed to be more complex. We use a lexicographic ordering of complexities, so that (m_1, n_1) is less complex than (m_2, n_2) if $m_1 < m_2$ or $m_1 = m_2, n_1 < n_2$. A factor model is called Pareto optimal if it satisfies the following two conditions: every less complex model has a strictly worse fit, and no equally complex model has strictly better fit. This means that the fit can not be improved without increasing the complexity, and that the complexity can not be reduced without detoriating the fit.

We characterize Pareto optimal models by optimizing the fit for a given bound on the complexity. This problem is analysed in three steps. In Section 4.1 we investigate two cases, that is, modelling with a specified behaviour and modelling with a restricted number of inputs where the number of states is left free. In Section 4.2 we derive Pareto optimal models of restricted complexity, where both the number of inputs and the number of states is limited. The optimality of models depends of course on the specification of the factor scheme, that is, on the choice of norms for the noise and on possible conditions of orthogonality and observability.

3.1 Optimal Models of Restricted Rank

First assume that the behaviour of the factor model has been specified a priori, so that the factor equations are given. The aim is to find a model with minimal error that satisfies these equations. Let $\mathcal{B}$ denote the given controllable linear system with polynomial representation $R(\sigma, \sigma^{-1})\,\hat{w} = 0$. The isometric image and kernel representations of the system are denoted respectively by $\hat{G}$ and $\tilde{G}$, so that $P_{\mathcal{B}} := \hat{G}\,\hat{G}^* = I - R^*(RR^*)^{-1}R$ is the projection operator onto the system and $\tilde{G}\,\tilde{G}^* = I - P_{\mathcal{B}}$ is the projection onto the set of behavioural equations. The following results hold true both for the mean squares and for the uniform norm.

Theorem 5 *Let w be a process with spectrum Σ and let $\mathcal{B}$ be the required behaviour of a factor model.*

(i) A latent process with optimal fit is given by $\hat{w}_0 := P_{\mathcal{B}} w$, with noise spectrum $\tilde{\Sigma}_0 = (I - P_{\mathcal{B}})\,\Sigma(I - P_{\mathcal{B}}) = \tilde{G}\,\tilde{G}^*\,\Sigma\,\tilde{G}\,\tilde{G}^*$. The corresponding factor model is observable, but in general not orthogonal.

(ii) Among orthogonal models, a latent process with optimal fit is given by $\hat{w}_0 := [I - \Sigma\,R^*(R\,\Sigma\,R^*)^{-1}R]w$, with corresponding noise spectrum $\tilde{\Sigma}_0 = \Sigma\,R^*(R\,\Sigma\,R^*)^{-1}R\,\Sigma = \Sigma\,\tilde{G}(\tilde{G}^*\,\Sigma\,\tilde{G})^{-1}\,\tilde{G}^*\,\Sigma$.

Proof.

(i) The relation $\tilde{G}^* \hat{w} = 0$ implies for the mean squares norm

$$\begin{array}{rcl} \| \tilde{\Sigma}^{1/2} \|_2^2 & = & \oint_{|z|=1} \mathrm{trace}(\hat{G}^* \tilde{\Sigma} \hat{G})(z) dz + \oint_{|z|=1} \mathrm{trace}(\tilde{G}^* \tilde{\Sigma} \tilde{G})(z) dz \\ & \geq & \oint_{|z|=1} \mathrm{trace}(\tilde{G}^* \Sigma \tilde{G})(z) dz. \end{array}$$

Therefore the misfit is minimal if and only if $\hat{G}^* \tilde{w} = 0$ holds so that $\hat{w} = (\hat{G}\hat{G}^* + \tilde{G}\tilde{G}^*)\hat{w} = \hat{G}\hat{G}^*(w - \tilde{w}) = P_{\mathcal{B}} w$. This model is also optimal for the uniform norm, since

$$\lambda_{\max}(\tilde{\Sigma}(z)) \geq \lambda_{\max}((\tilde{G}\tilde{G}^* \tilde{\Sigma} \tilde{G}\tilde{G}^*)(z)) \quad = \quad \lambda_{\max}((\tilde{G}\tilde{G}^* \Sigma \tilde{G}\tilde{G}^*)(z))$$

holds for all points z of the unit circle. This optimal model is, in general, not orthogonal since $P_{\mathcal{B}} \Sigma (I - P_{\mathcal{B}})$ is not zero in general.
(ii) We show that $\tilde{\Sigma}(z) \geq \tilde{\Sigma}_0(z)$ holds for all points z of the unit circle. For simplicity of notation we omit the argument z in the following. Let $G = [\hat{G}, \tilde{G}]$, then because of $\tilde{G}^* \hat{\Sigma} = 0$ it follows that $\tilde{G}^* \Sigma = \tilde{G}^* \tilde{\Sigma}$ and hence

$$\begin{array}{rcl} G^* \tilde{\Sigma} G & = & \begin{pmatrix} \hat{G}^* \tilde{\Sigma} \hat{G} & \hat{G}^* \Sigma \tilde{G} \\ \tilde{G}^* \Sigma \hat{G} & \tilde{G}^* \Sigma \tilde{G} \end{pmatrix} \\ & \geq & \begin{pmatrix} (\hat{G}^* \Sigma \tilde{G})(\tilde{G}^* \Sigma \tilde{G})^{-1}(\tilde{G}^* \Sigma \hat{G}) & \hat{G}^* \Sigma \tilde{G} \\ \tilde{G}^* \Sigma \hat{G} & \tilde{G}^* \Sigma \tilde{G} \end{pmatrix} \\ & = & G^* \Sigma \tilde{G}(\tilde{G}^* \Sigma \tilde{G})^{-1} \tilde{G}^* \Sigma G \end{array}$$

The above inequality is a consequence of the fact that $G^* \tilde{\Sigma} G \geq 0$. So all orthogonal factor models with behaviour $\mathcal{B}$ must satisfy $\tilde{\Sigma} \geq \Sigma \tilde{G}(\tilde{G}^* \Sigma \tilde{G})^{-1} \tilde{G}^* \Sigma = \tilde{\Sigma}_0$. This shows the second expression for $\tilde{\Sigma}_0$. The first expression follows from the fact that $\tilde{G} = R^* Q$ where Q is a spectral factor of $(RR^*)^{-1}$, that is $QQ^* = (RR^*)^{-1}$. □

The optimal factor model is unique in case of the mean squares norm but in general not in case of the uniform norm. If we are interested in factor behaviours only, then the above results show that we may restrict the attention to observable models. This leaves four factor schemes of interest, that is, for the mean squares and the uniform norm and according to whether orthogonality is imposed or not. We define the distance between a behaviour and a spectral density as the fit of the optimal factor model with this behaviour. That is, the misfit function is given by

$$\mathrm{d}(\Sigma, \mathcal{B}) = \| \tilde{\Sigma}_0^{1/2} \| \tag{14}$$

where $\tilde{\Sigma}_0$ is the noise spectrum of the optimal factor models for $\mathcal{B}$ given in Theorem 5 and where $\tilde{\Sigma}_0^{1/2}$ denotes a spectral factor of $\tilde{\Sigma}_0$. We use the same notation for the four different factor schemes.

Next we describe optimal models of restricted rank, so that only the number of inputs of the latent process is restricted but not the number of state variables. Under the assumptions A1-A4 of Section 2.2, the observed spectrum Σ is a well-defined matrix function on the unit circle that can be pointwise decomposed in terms of its eigenvalues and eigenvectors as $\Sigma = U\Lambda U^*$. Here U is a $q \times q$ unitary matrix function, i.e., $UU^* = U^*U = I_q$, and Λ is a diagonal matrix of ordered eigenfunctions. For simplicity we assume that the eigenvalues are distinct everywhere on the unit circle, so that $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_q)$ with $\lambda_1(z) > \lambda_2(z) > \cdots > \lambda_q(z) > 0$ on the unit circle. Let $U = [U_1, U_2]$, where U_1 consists of the first m columns of U and U_2 of the remaining columns, and let $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$ be a corresponding partitioning. The principal component model of rank m is defined by the factor $\hat{w} = U_1U_1^*w$ and noise $\tilde{w} = U_2U_2^*w$. In terms of the spectra this gives

$$\hat{\Sigma}_m = U_1\Lambda_1U_1^*, \quad \tilde{\Sigma}_m = U_2\Lambda_2U_2^*, \quad \Sigma_c = 0 \tag{15}$$

Under the above assumptions, this model is well-defined and unique, see [3, theorems 9.3.1, 9.3.2 and 9.3.3], and it is clearly observable and orthogonal. The latent process spectrum has rank m, but the factor behaviour will in general be trivial, that is, it will be $(\mathbf{R}^q)^{\mathbf{Z}}$. This is because in general there exist no nontrivial polynomial equations such that $R(z, z^{-1})\,\hat{\Sigma}_m(z) = 0$.

The following result states that the principal component model has optimal fit, and that it can be approximated arbitrarily closely by factor models with complexity (m, n) if the number of state variables n is chosen sufficiently large. The results hold true for all factor schemes, that is, for mean squares and uniform fit and irrespective whether orthogonality and observability are imposed or not.

Theorem 6

(i) No factor model of complexity (m, n) has better fit than the fit $\| \tilde{\Sigma}_m^{1/2} \|$ of the principal component model of rank m.

(ii) For every $\varepsilon > 0$ there is a factor model of complexity (m, n), for some finite n, with better fit than $\| \tilde{\Sigma}_m^{1/2} \| + \varepsilon$.

Proof. Under the assumption $\lambda_1(z) > \lambda_2(z) > \cdots > \lambda_q(z) > 0$ for all $|z| = 1$, the eigenvector matrix $U(z)$ has an absolutely summable Laurent series expansion, see [3, theorems 9.3.1, 9.3.2 and 9.3.3]. This implies that $\tilde{w}_m = U_2U_2^*w$ and $\hat{w}_m = w - \tilde{w}_m = U_1U_1^*w$ are well-defined processes.

(i) As $\Sigma(z)$ is continuous on the unit circle it follows that also the eigenvalues $\lambda_i(z)$ are continuous functions, see Lemma 20 in the appendix, and thus $\| \tilde{\Sigma}_m^{1/2} \|_2^2 = \oint_{|z|=1}\{\lambda_{m+1}(z) + \cdots + \lambda_q(z)\}dz$ and $\| \tilde{\Sigma}_m^{1/2} \|_\infty^2 = \sup_{|z|=1} \lambda_{m+1}(z)$ are well-defined. If $\tilde{G}$ is the isometric kernel representation of a behaviour $\mathcal{B}$, then the optimal noise covariance corresponding to $\mathcal{B}$ is according to Theorem 5 given by $\tilde{\Sigma} = \tilde{G}\tilde{G}^*\,\Sigma\,\tilde{G}\tilde{G}^*$. As $\tilde{G}^*(z)\,\tilde{G}(z) = I$, Lemma 20 implies the optimality of the principal component model.

(ii) Since $U_2(\sigma)$ is an absolutely summable filter, we can find a positive integer N and a finite filter $\tilde{G}_N(\sigma) = \sum_{|k|\leq N} \tilde{G}_k \, \sigma^k$ such that $\|U_2 - \tilde{G}_N \|_\infty$ is arbitrarily small. Thus we can choose N such that $\|I - \tilde{G}_N^* \, \tilde{G}_N \|_\infty$ and also $\|U_2 U_2^* - \tilde{G}_N(\tilde{G}_N^* \, \tilde{G}_N)^{-1} \, \tilde{G}_N^* \|_\infty$ become arbitarily small. The transferfunction $P_N = \tilde{G}_N(\tilde{G}_N^* \, \tilde{G}_N)^{-1} \, \tilde{G}_N^*$ is a rational projection matrix of rank m, so that $(I - P_N)$ is the isometric image representation of a behaviour $\mathcal{B}_N$ with m inputs and a finite number of states. Then analogous to the proof of Proposition 22 in the appendix it follows that $\| \tilde{\Sigma}_N - \tilde{\Sigma}_m \|_\infty \to 0$ and thus $\| \tilde{\Sigma}_N^{1/2} \| \to \| \tilde{\Sigma}_m^{1/2} \|$ by Lemma 21. Here U_2 corresponds to $\tilde{G}_0$ in the proof of Proposition 22 and this proof can easily be extended to the case where $\tilde{G}_0 = U_2$ is not rational but only absolutely summable. □

So the principal component model gives an optimal reduced rank approximation of the spectrum. Further this gives a first idea of achievable combinations of complexity and fit. A sufficient condition for the existence of a factor model with fit δ and complexity (m, n), for some finite n, is that $\| \tilde{\Sigma}_m^{1/2} \| < \delta$, and a necessary condition is that $\| \tilde{\Sigma}_m^{1/2} \| \leq \delta$.

We conclude this section by considering the effect of using weighted norms, or stated otherwise, the effect of prefiltering the observed process. Let Q be a $q \times q$ positive definite matrix function which is bounded on the unit circle. Then the Q-weighted norm is defined as $\| \tilde{w} \|_Q = \|T^* \, \tilde{w} \|$ for a spectral factorization $Q = TT^*$. This norm is well-defined, as it does not depend on the choice of the spectral factor.

Proposition 7 *Let $\mathcal{B}$ be a controllable linear system of complexity (m, n). Then there is a choice of Q-weights such that $\mathcal{B}$ is the behaviour of a factor model that minimizes the Q-weighted norm over the set of all factor models with m inputs.*

Proof. Let $R(\sigma, \sigma^{-1})$ be a full row rank polynomial matrix with rows that form a basis for the set of laws of the behaviour $\mathcal{B}$. As $\| \tilde{w} \|_Q = \|T^* \, \tilde{w} \|$ we can use the result of Theorem 6 on the transformed data $\bar{w} := T^* w$, with spectrum $T^* \, \Sigma \, T$. The transformed latent process $\hat{\bar{w}} = T^* \, \hat{w}$ satisfies the relation $R(T^*)^{-1} \hat{\bar{w}} = 0$. Thus by Theorem 6, $\mathcal{B}$ is optimal with respect to the weighted norm $\| \tilde{w} \|_Q$ if $R(T^*)^{-1}$ is a basis of the left eigenspace of $T^* \, \Sigma \, T$ corresponding to $q - m$ smallest eigenvalues, pointwise on the unit circle. In this case $\bar{w} = \hat{\bar{w}} + \tilde{\bar{w}}$ is the princial component model for the transformed data.

Now let $\bar{S}(\sigma, \sigma^{-1})$ be a full column rank polynomial matrix with columns that form a basis of the right kernel of R, i.e. $R\bar{S} = 0$, and let $S = \Sigma^{-1} \, \bar{S}$ and $Q = \Sigma^{-1} + SS^*$. If $\bar{Q} = \Sigma^{*/2} \, Q \, \Sigma^{1/2}$, then it follows that $R \, \Sigma^{1/2} \, \bar{Q} = R \, \Sigma^{1/2}$ and $S^* \, \Sigma^{1/2} \, \bar{Q} = (I + S^* \, \Sigma \, S) S^* \, \Sigma^{1/2}$. Thus the $q - m$ smallest eigenvalues of $\bar{Q}(z)$ are equal to 1 and $R(z, z^{-1}) \, \Sigma^{1/2}(z)$ is a basis of the corresponding left eigenspace. Let $Q = TT^*$ and $\bar{\Sigma} = T^* \, \Sigma \, T$, then there holds $x \, \Sigma^{1/2} \, \bar{Q} = \lambda x \, \Sigma^{1/2}$ if and only if $x(T^*)^{-1} \bar{\Sigma} = \lambda x (T^*)^{-1}$. So the $q - m$ smallest eigenvalues of

$\bar{\Sigma} = T^* \Sigma T$ are equal to one and $U_2 = R(T^*)^{-1}$ is a basis of the corresponding eigenspace. This shows that T is the appropriate transformation and Q the appropriate norm. □

This shows that the choice of norms is decisive for the obtained behaviours. So in practical applications it is imperative to take care of appropriate weighting of the data. In our opinion the norms should not be chosen on mathematical grounds alone but have to be related to the information and objectives of each specific application. Here we will further restrict attention to the unweighted norms, which may be relevant in applications if the observed variables have been transformed appropriately.

3.2 Optimal Models of Restricted Complexity

A straightforward method for determining Pareto optimal models is to fix the complexity and to optimize the fit under this constraint. A model of optimal fit is then Pareto optimal if there are no less complex models of at least equal fit. For complexity (m, n) this can be checked by comparing, first, with the optimal fit of models of complexity $(m, n-1)$ and, second, with the fit achievable by models having less than m inputs. The second comparison is simplified by the result of Theorem 6 for the principal component model of rank $m-1$. Because of these considerations, we restrict our attention to the determination of optimally fitting models of given complexity.

The main complication of the corresponding optimization problem is that the set of systems of given complexity (m, n) is not convex and also not compact. We restrict the attention to the mean squares norm and consider both the factor schemes with and without orthogonality. We will not investigate several other questions that are of interest in this context, such as the existence and unicity of optimal models and the case of the uniform norm.

The solution for the mean squares norm is given in terms of the so-called global total least squares algorithm presented in [21]. Let $W = (W_1, \cdots, W_r)$ be a square summable $q \times r$ matrix sequence, that is, with $\|W\|_2^2 := \sum_{t=-\infty}^{\infty} \|W(t)\|_2^2 < \infty$ where $\|W(t)\|_2$ denotes the Frobenius norm of the matrix $W(t)$. Further let the l_2-distance between this sequence and a linear system $\mathcal{B}$ be defined as $\mathrm{d}(W, \mathcal{B}) := \min\{\|W - V\|_2; V = (V_1, \cdots, V_r) \text{ with } V_i \in \mathcal{B}, i = 1, \cdots, r\}$. The objective in global total least squares is to determine an optimal model of restricted complexity, that is, which minimizes the l_2-distance over the set of controllable systems with m inputs and n states. In general the optimal model exists and is unique, but existence and uniqueness may fail to hold true in exceptional cases. For algorithmic details we refer to [21] and [20] where a Gauss-Newton algorithm for the involved projections is described. If $\mathcal{B}$ is the optimal system, then $P_{\mathcal{B}} W$ is called the optimal l_2-approximation of W.

Theorem 8 *Let $w = T\varepsilon$ be a given process with spectrum $\Sigma = TT^*$. For given complexity (m,n), a factor models with optimal mean squares fit is given by $w = \hat{w} + \tilde{w}$, where $\hat{w} = \hat{T}\varepsilon$ and $\tilde{w} = (T - \hat{T})\varepsilon$. Here $\hat{T}$ is the optimal l_2-approximation of complexity (m,n) for the spectral factor T. This model is observable, but in general not orthogonal.*

Proof. According to Theorem 5, it is no restriction of generality if we consider only observable models. So let $\hat{w}(t) = [G_t(\sigma, \sigma^{-1})\varepsilon](t)$, then assumption A1 of joint stationarity of w and $\hat{w}$ implies that G_t is time invariant, say $G_t = G$. This means that we can write $\hat{w} = F(\sigma)w = G(\sigma)\varepsilon$ for some transfer function $G(\sigma) = F(\sigma)T(\sigma)$. As ε has full rank, it follows that the latent process has complexity (m,n), that is, $R\hat{w} = 0$ for a polynomial matrix R representing a system $\mathcal{B}$ with complexity (m,n), if and only if $RG = 0$, that is, all columns of G should belong to the system $\mathcal{B}$. The noise $\tilde{w} = (T-G)\varepsilon$ has spectrum $(T-G)(T-G)^*$ and mean squares norm $\| \tilde{w} \|_2^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \text{trace}\{(T-G)(T-G)^*(e^{-i\lambda})\}d\lambda$. But this is precisely equal to $\|T - G\|_2^2$, the l_2-distance between T and G. So this minimization problem is the l_2- approximation problem for T where each of the q columns of G should belong to the same system of complexity (m,n). The optimal choice over this class is by definition given by $\hat{T}$.

It can be shown that the factor filter $\hat{T}$ and the noise filter $\tilde{T} := T - \hat{T}$ satisfy $\hat{T}^* \tilde{T} = 0$, but in general $\Sigma_c = \hat{T}\tilde{T}^* \neq 0$ so that the processes $\hat{w}$ and $\tilde{w}$ are not orthogonal. □

Next we characterize optimal models under the condition of orthogonality. In order to simplify the analysis we restrict the attention to observed processes with rational spectrum Σ and use the alternative definition of complexity in terms of the effective noise space, see Definition 3.

Theorem 9 *Let $w = T\varepsilon$ be a given process with spectrum $\Sigma = TT^*$. For given noise complexity (m,n), an orthogonal factor model with optimal mean squares fit is given by $w = \hat{w} + \tilde{w}$, where $\hat{w} = S\varepsilon$ and $\tilde{w} = (T - S)\varepsilon$. Here S^* is the optimal l_2-approximation of complexity (m,n) for the adjoint T^* of the spectral factor T.*

Proof. Within this setting a latent process is given by $\hat{w} = TF\varepsilon$ where the factor noise space $\mathcal{N} = \text{im}(F)$ has complexity (m,n). The noise is then given by $\tilde{w} = T(I - F)\varepsilon$, and the orthogonality condition is equivalent to requiring $TF(I-F)^*T^* = 0$. As T has full rank everywhere, it follows that $F = F^* = F^2$ is a projection, namely the orthogonal projection onto the system $\mathcal{N}$. The noise has norm $\| \tilde{w} \|_2^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \text{trace}\{T(I-F)T^*(e^{-i\lambda})\}d\lambda$. This is equal to $\|(I - F)T^*\|_2^2 = \|T^* - S^*\|_2^2$, where each column of S^* is the optimal l_2-approximation within the system $\mathcal{N}$ of the corresponding column of T^*, as F is the projection onto this system. The optimal choice of the model, that is, of F or equivalently of $\mathcal{N}$ of complexity (m,n), is precisely the optimal l_2-approximation problem of T^*. □

3.3 Illustrations

3.3.1 The Static Case

The foregoing results can easily be applied for the case of static factor models. Let w be a white noise process, so that the spectrum Σ is a constant function, the covariance matrix of the process. The principal component model is then obtained by the eigenvalue decomposition of the matrix Σ. The optimal latent process with m factors is given by the projection of the observations onto the space spanned by the eigenvectors corresponding to the m leading eigenvalues of Σ. Therefore, in the optimal factor model both the latent process and the noise are white noise processes. It follows from Theorem 6(i) that the principal component model is Pareto optimal among all models of complexity (m,n) for all $n \geq 0$. That is, no gain of fit is possible by allowing for dynamic equations.

For the static case, the result in Proposition 7 has also been pointed out in [15] and [16]. In the ordinary least squares scheme the indeterminateness of optimal models is resolved by the assumption that certain variables are noise free, i.e., that a principal submatrix of the noise covariance $\tilde{\Sigma}$ is zero. In terms of the weighting matrix Q this means that certain noise directions are assigned an infinite weight. In our approach, however, we treat all variables in a symmetric way.

3.3.2 Dynamic System Example

Next we consider the dynamic errors in variables system described in Section 2.4.2, and we use the notation introduced there. So let the spectrum Σ be given, and assume that the complexity (m,n) has been specified with $m = 1$ and $n \geq d$. The principal component decomposition for fixed frequency is easily obtained, with eigenvalues $\lambda_1 = 2 + gg^*$ and $\lambda_2 = 1$ and the eigenvector corresponding to λ_2 given by $(-g, 1)^*$. We denote the corresponding latent process by $\hat{w}_* = (\hat{u}_*, \hat{y}_*)$ and the noise process by $\tilde{w}_*$. This shows that the principal component model has a behaviour that is finite dimensional, and this model is Pareto optimal among all models of complexity $(1,n)$ with $n \geq d$. The underlying transfer function g has been identified, because $\hat{y}_* = g\hat{u}_*$.

Although the underlying behaviour has been identified, this is not the case for the true latent process and noise process. This can be seen from the spectral properties of the noise processes. The noise that affects the data has spectrum I_2 of rank 2, whereas the noise $\tilde{w}_*$ has a spectrum of only rank 1. Further the factor model $w = \hat{w} + \tilde{w}$ has a mean squares error $\| \tilde{w} \|_2 = \sqrt{2}$ whereas the principal component model has error $\| \tilde{w}_* \|_2 = 1$. We remark that both models are in fact optimal for the uniform norm.

This shows that in this case the Pareto optimal model indeed identifies the latent transfer function g from the observed spectrum Σ, at least when the complexity is not chosen too small. We should remark that this result depends in a crucial way on our assumptions on the way the data are generated. For example, if the observation noise $\tilde{w}$ would not be white then Pareto optimal models will in general not have transfer function g. In terms of Proposition 7 this would require an appropriate prefiltering of the data. In our example, the required filter Q is the identity, that is, our data generating process is such that the unweighted norm is appropriate to identify the underlying transfer function. For practical applications this means that, in order to find good approximations of the underlying system, one should incorporate available information on the noise properties.

4 Consistency

4.1 System Topology

We introduce the topologies on linear systems and spectra that we will use in our analysis of continuity properties of factor models. For linear systems the gap metric is defined in terms of the projections described at the end of Section 2.1.

Definition 5 *Let $\mathcal{B}_1, \mathcal{B}_2$ be linear systems with isometric image representations $\hat{G}_1$ and $\hat{G}_2$ respectively, then the gap between these systems is defined by*

$$\mathrm{d}(\mathcal{B}_1, \mathcal{B}_2) = \| \hat{G}_1 \hat{G}_1^* - \hat{G}_2 \hat{G}_2^* \|_\infty \tag{16}$$

This corresponds to the usual definition of the gap between two closed linear subspaces of a Hilbert space as $\|P_1 - P_2\|$, where P_1 and P_2 are the orthogonal projection operators onto the two spaces. Here $\hat{G}_i \hat{G}_i^*$ is the orthogonal projection onto the set of square summable time series in the behaviour $\mathcal{B}_i$, $i = 1, 2$.

Proposition 10

(i) The gap d *is a metric on the class of controllable linear systems.*

(ii) In terms of system restrictions, if $\tilde{G}_i$ denotes an isometric kernel representation of $\mathcal{B}_i$, $i = 1, 2$, then $\mathrm{d}(\mathcal{B}_1, \mathcal{B}_2) = \| \tilde{G}_1 \tilde{G}_1^* - \tilde{G}_2 \tilde{G}_2^* \|_\infty$.

(iii) If two systems have a different number of inputs, then their gap equals one.

Proof. (i) This holds true for so-called l_2 systems, and this implies the same result for controllable systems as these are in one-to-one correspondence with l_2 systems. See corollaries 3-4 and 5-3 of chapter 4 in [14].

(ii) This follows from the fact that $[\hat{G}, \tilde{G}]$ is inner, so that $\hat{G} \hat{G}^* + \tilde{G} \tilde{G}^* = I$.

(iii) See Proposition 5-5 of chapter 4 in [14]. □

In the following we denote by $\mathbf{B}(m, n)$ the set of all controllable linear systems with m inputs and n states, by $\overline{\mathbf{B}}(m, n) := \bigcup_{k=1}^{n} \mathbf{B}(m, k)$ the set of all controllable linear systems with m inputs and at most n states, and by $\mathbf{B} := \bigcup_{m=0}^{q} \bigcup_{n=0}^{\infty} \mathbf{B}(m, n)$ the set of all controllable linear systems.

Proposition 11

(i) For $n > 0$ the set $\mathbf{B}(m, n)$ is neither open nor closed in $\mathbf{B}$.

(ii) The set $\overline{\mathbf{B}}(m, n)$ is the closure of $\mathbf{B}(m, n)$ in $\mathbf{B}$.

(iii) The sets $\mathbf{B}$ *and* $\overline{\mathbf{B}}(m,n)$*, for* $n > 0$*, are not compact.*

Proof. (i) For $n = 0$ the only controllable systems are described by the isometric state parameters $(A, B, C, D) = (-, -, -, D)$ with corresponding static projection operator DD'. It follows that $\mathbf{B}(m, 0)$ is a compact set, and this will be described in more detail in Section 4.4. We will now consider the case $n > 0$.

In order to show that $\mathbf{B}(m,n)$ is not open it suffices to construct a sequence of systems $\mathcal{B}_k \in \mathbf{B}(m, n+1)$ with $\mathrm{d}(\mathcal{B}_k, \mathcal{B}_0) \to 0$ where $\mathcal{B}_0 \in \mathbf{B}(m,n)$. Let (A_0, B_0, C_0, D_0) be a minimal isometric state representation of $\mathcal{B}_0$ and let $a \in \mathbf{R}, b \in \mathbf{R}^{1\times m}$ and $c \in \mathbf{R}^{q\times 1}$ be such that $A = \begin{pmatrix} A_0 & 0 \\ 0 & a \end{pmatrix}$, $B = \begin{pmatrix} B_0 \\ b \end{pmatrix}$, $C_k = (C_0, \varepsilon_k c)$ is an observable and contollable quadruple for all $\varepsilon_k > 0$. The system $\mathcal{B}_0$ has transfer function $\hat{G}_0 = D_0 + C_0(zI - A_0)^{-1}B_0$, and let the system $\mathcal{B}_k$ be defined by the transfer function $\hat{G}_k = D_0 + C_k(zI - A)^{-1}B = \hat{G}_0 + \varepsilon_k(z-a)^{-1}cb$ with $\varepsilon_k \to 0$ for $k \to \infty$. Then $\mathcal{B}_k \in \mathbf{B}(m, n+1)$ and clearly $\|G_k - G_0\|_\infty \to 0$ and also $\mathrm{d}(\mathcal{B}_k, \mathcal{B}_0) = \| \hat{G}_k(\hat{G}_k^* \hat{G}_k)^{-1} \hat{G}_k^* - \hat{G}_0 \hat{G}_0^* \|_\infty \to 0$ for $k \to \infty$.

That $\mathbf{B}(m,n)$ is not closed follows in a similar way by constructing a sequence in $\mathbf{B}(m,n)$ that converges to a system in $\mathbf{B}(m, n-1)$.

(ii) Let $cl\,\mathbf{B}(m,n)$ denote the closure of $\mathbf{B}(m,n)$. Systems with $m' \neq m$ do not belong to this closure, as such systems have gap one with respect to all systems in $\mathbf{B}(m,n)$, see Proposition 10(iii). Systems with m inputs and less than n states can be obtained as the limit of sequences of systems in $\mathbf{B}(m,n)$, by similar constructions as in the proof of (i). It remains to prove that systems in $\mathbf{B}(m,n')$ with $n' > n$ do not belong to $cl\,\mathbf{B}(m,n)$. Let $\mathcal{B} \in \mathbf{B}(m,n')$ with $n' > n$ have isometric image representation $\hat{G}$, then the projection operator $P = \hat{G}\hat{G}^*$ is a rational function with rank m and McMillan degree $2n'$. As projection operators corresponding to systems in $\mathbf{B}(m,n)$ have rank m and McMillan degree $2n$, it follows that such operators can not converge to P, so that $\mathcal{B}$ does not belong to $cl\,\mathbf{B}(m,n)$.

(iii) As $\mathbf{B}$ is a metric space, it suffices to prove that there exists a sequence of systems $\mathcal{B}_k \in \overline{\mathbf{B}}(m,n)$ which has no convergent subsequence in the set $\mathbf{B}$ of all controllable linear systems. Consider the case $q = 2, m = 1, n = 1$ and the systems described by the isometric state parameters $\begin{pmatrix} a & \beta\delta C'D \\ \gamma C & \delta D \end{pmatrix}$, where $0 < a < 1$ is a real number, C and D are 2×1 vectors of unit length, and β, γ, δ are real numbers to obtain an isometric matrix, that is, $\gamma = \sqrt{1-a^2}$, $\beta = -\gamma/a$ and $\delta = \{1 + \beta^2(C'D)^2\}^{-1/2}$. To guarantee minimality it is further assumed that $C'D \neq 0$. The corresponding isometric image representations are given by $\hat{G}(z) = \delta D + \beta\gamma\delta(C'D)C(z-a)^{-1}$, and the projection operators by $P = \hat{G}\hat{G}^*$. If $a \uparrow 1$ then $\gamma \to 0, \beta \to 0$ and $\delta \to 1$, so that the pointwise limit of $\hat{G}(z)$ is D for $z \neq 1$ and $\hat{G}(1)$ converges to $D - 2(C'D)C$. If the corresponding sequence of systems would have a limiting point, say with projection operator P_0, then it should hold that $\|P_0 - P\|_\infty \to 0$ for $a \uparrow 1$. As $P_0(z)$ is continuous on the unit circle the only candidate for P_0 is given by DD', but as $P(z)$ is also continuous and $P(1) \not\to DD'$ for $a \uparrow 1$ it follows that no subsequence can converge to a system in $\mathbf{B}$. □

In our analysis not only the distance between two systems, but also the distance between two sets of systems is of relevance. If $\mathbf{B}_1$ and $\mathbf{B}_2$ are two compact subsets of $\mathbf{B}$, then the Hausdorff distance between these sets is defined as

$$\mathrm{d_H}(\mathbf{B}_1, \mathbf{B}_2) := \max\{\rho(\mathbf{B}_1, \mathbf{B}_2), \rho(\mathbf{B}_2, \mathbf{B}_1)\}. \tag{17}$$

where $\rho(\mathbf{B}_1, \mathbf{B}_2) := \sup_{\mathcal{B}_1 \in \mathbf{B}_1} \inf_{\mathcal{B}_2 \in \mathbf{B}_2} \mathrm{d}(\mathcal{B}_1, \mathcal{B}_2)$.

In order to investigate continuity properties we also need a topology on the set of spectral densities. We use the metric defined by

$$\mathrm{d}(\Sigma_1, \Sigma_2) = \| \Sigma_1 - \Sigma_2 \|_\infty := \sup_{\lambda \in [-\pi, \pi]} \lambda_{\max}\{\Sigma_1(e^{-i\lambda}) - \Sigma_2(e^{-i\lambda})\} \tag{18}$$

Under Assumption A4 the spectra are bounded on the unit circle, so that this is a well-defined metric.

4.2 Continuity

We consider the relation between observed spectra and identified factor behaviours. For given spectrum Σ, complexity (m, n) and noise bound δ, we denote by $\mathbf{B}(\Sigma; \delta, m, n) \subseteq \mathbf{B}(m, n)$ the set of all behaviours of factor models $w = \hat{w} + \tilde{w}$ satisfying the conditions that the factor behaviour has m inputs and n states and that the noise process has norm $\| \tilde{w} \| \leq \delta$. So this corresponds to the factor scheme with bounded noise. The set $\mathbf{B}(\Sigma; \delta, m, n)$ depends of course on the measure of fit and on the possible condition of orthogonality. As the results in this section hold true for all the four corresponding factor schemes, we will make no explicit distinction between them. Systems in $\mathbf{B}(\Sigma; \delta, m, n)$ are called feasible for the data Σ and the specified complexity and fit. The feasibility of a given behaviour can be checked by means of the results in Theorem 5.

Proposition 12

(i) The set of feasible systems $\mathbf{B}(\Sigma; \delta, m, n)$ depends on whether orthogonality is imposed or not, but it does not depend on whether observability is imposed or not.

(ii) The set $\mathbf{B}(\Sigma; \delta, m, n)$ is closed in $\mathbf{B}(m, n)$, but in general not in $\mathbf{B}$.

(iii) If $\mathcal{B} \in \mathbf{B}(\Sigma; \delta, m, n)$ has fit strictly better than δ, then it is an inner point of $\mathbf{B}(\Sigma; \delta, m, n)$.

Proof. (i) This follows from Theorem 5.

(ii) The set $\mathbf{B}(\Sigma; \delta, m, n)$ is closed in $\mathbf{B}(m, n)$ by Proposition 22 in the appendix, but not in $\mathbf{B}$ as follows from Proposition 11(i).

(iii) This is immediate from Proposition 22. □

In order to use the Hausdorff metric (17) we next formulate a sufficient condition for compactness. We call a state dimension n *minimal* for given (Σ, δ, m) if there exists a feasible model of complexity (m, n) but not one of complexity (m, n') with $n' < n$, that is, if $\mathbf{B}(\Sigma; \delta, m, n) \neq \emptyset$ and $\mathbf{B}(\Sigma; \delta, m, n') = \emptyset$ for all $n' < n$. If we are only interested in Pareto optimal models, then this minimality condition can be imposed without loss of generality.

Proposition 13 *If n is minimal for (Σ, δ, m) then the set of feasible systems $\mathbf{B}(\Sigma; \delta, m, n)$ is compact.*

Proof. We prove this in terms of isometric state space representations. For this purpose we first describe this parametrization in some more detail. By definition, systems in $\mathbf{B}(m, n)$ are controllable and so can be represented by an isometric state model that satisfies (5). Let $\Pi(m, n) \subset \mathbf{R}^{(n+q)\times(n+m)}$ be the set of all such minimal isometric system matrices and let $\Pi = \bigcup_{m=0}^{q} \bigcup_{n=1}^{\infty} \Pi(m, n)$. On this set we define the metric $\mathrm{d}(\pi_1, \pi_2) = \|\pi_1 - \pi_2\|_\infty$ if $(m_1, n_1) = (m_2, n_2)$ and $\mathrm{d}(\pi_1, \pi_2) = 3$ otherwise. It is easily verified that this is a metric on Π and that $\Pi(m, n)$ is open in Π. That the parametrization of $\mathbf{B}$ by Π is continuous can be seen as follows. Let $\pi_k \to \pi_0$, then for k sufficiently large there holds $(m_k, n_k) = (m_0, n_0)$. As π_0 is a minimal isometric representation it follows that $A_0 A_0' + C_0 C_0' = I$ with (A_0, C_0) observable, so that A_0 has all its eigenvalues strictly within the unit circle. Then the mapping from (A, B, C, D) to the isometric image representation $\hat{G} = D + C(zI - A)^{-1}B$ is continuous in π_0, and so $\mathrm{d}(\mathcal{B}_k, \mathcal{B}_0) = \|\, \hat{G}_k \hat{G}_k^* - \hat{G}_0 \hat{G}_0^* \,\| \to 0$ for $k \to \infty$.

Because the parametrization is continuous, in order to prove that $\mathbf{B}(\Sigma; \delta, m, n)$ is a compact subset of $\mathbf{B}$ it suffices to prove that the corresponding set of parameters denoted by $\Pi_0 \subset \Pi$ is compact. As $\Pi(m, n) \subset \Pi$ is open it suffices to prove that Π_0 is a compact subset of $\Pi(m, n)$, or also that it is a closed and bounded subset of the Euclidean space $\mathbf{R}^{(n+q)\times(n+m)}$. Because of the isometry condition boundedness is evident, so that it remains to prove the closedness of Π_0. We prove this by contradiction.

So suppose that there is a sequence of systems $\mathcal{B}_k \in \mathbf{B}(\Sigma; \delta, m, n)$ with minimal isometric represenations $(A_k, B_k, C_k, D_k) \to (A_0, B_0, C_0, D_0)$ so that the system $\mathcal{B}_0$ corresponding to these limit parameters does not belong to $\mathbf{B}(\Sigma; \delta, m, n)$. Then A_0 has eigenvalues on the unit circle. Indeed, if this were not the case then the parametrization would be continuous in (A_0, B_0, C_0, D_0) and hence, by Proposition 22, it would follow that $\mathrm{d}(\Sigma, \mathcal{B}_0) = \lim \mathrm{d}(\Sigma, \mathcal{B}_k) \leq \delta$. As $\mathcal{B}_0$ has m inputs and n is assumed to be minimal for (Σ, δ, m) it would follow that $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta, m, n)$, contradicting our assumption. Now state directions corresponding to eigenvectors of unit eigenvalues of A_0 are not observable, because of the isometry condition $A_0' A_0 + C_0' C_0 = I$. So the state space for $\mathcal{B}_0$ can be reduced by deleting such unobservable directions. Let (A, B, C, D_0) be the restriction of (A_0, B_0, C_0, D_0) to the observable subspace, so that A has all its eigenvalues strictly within the unit circle. Because the two representations describe the same system $\mathcal{B}_0$ with the same driving variables, it follows that $G_0(z) := D_0 + C(zI - A)^{-1}B = D_0 + C_0(zI - A_0)^{-1}B_0$ pointwise on the unit circle, with the exception of the eigenvalues $\{e^{-i\lambda_j}; j = 1, \cdots, r\}$ of A_0. Moreover, as G_0 is the pointwise limit of $G_k = D_k + C_k(zI - A_k)^{-1}B_k$ it follows that G_0 is an isometric image representation of $\mathcal{B}_0$, with m inputs and at most $n - r$ states.

We consider first the factor scheme without orthogonality and with the uniform norm. Using the notation (14), we obtain from Theorem 5(i) that the fit of the system in this case is given by $\mathrm{d}(\Sigma, \mathcal{B}_0) = \| \tilde{\Sigma}_0^{1/2} \|_\infty$ where $\tilde{\Sigma}_0 = (I - G_0G_0^*)\,\Sigma(I - G_0G_0^*)$. As n is minimal for (Σ, δ, m) and $\mathcal{B}_0$ has less than n states, it follows that $\sup_{\lambda \in [-\pi,\pi]} \lambda_{\max}\{\tilde{\Sigma}_0(e^{-i\lambda})\} > \delta^2$. Because of the continuity of $G_0(z)$ and $\Sigma(z)$ on the unit circle there exists an $\varepsilon > 0$ so that also $\sup_{\lambda \in \Lambda} \lambda_{\max}\{\tilde{\Sigma}_0(e^{-i\lambda})\} > \delta^2$ where $\Lambda = \{\lambda \in [-\pi, \pi]; |\lambda - \lambda_j| \geq \varepsilon \text{ for all } j = 1, \cdots, r\}$. As G_k converges pointwise to G_0 on the compact set Λ this implies that for k sufficiently large $\{\mathrm{d}(\Sigma, \mathcal{B}_k)\}^2 \geq \sup_{\lambda \in \Lambda} \lambda_{\max}\{(I - G_kG_k^*)\,\Sigma(I - G_kG_k^*)(e^{-i\lambda})\} > \delta^2$, but this contradicts that $\mathcal{B}_k \in \mathbf{B}(\Sigma; \delta, m, n)$. This proves compactness for the factor scheme without orthogonality and with the uniform norm.

The result for the orthogonal factor scheme with uniform norm follows in a similar way by using Theorem 5(ii). For the mean squares norm the reasoning is similar. Under the assumptions as before there would exist an $\varepsilon > 0$ such that $\frac{1}{2\pi}\int_\Lambda \mathrm{trace}\{\tilde{\Sigma}_0(e^{-i\lambda})\}d\lambda > \delta^2$, and as G_k converges uniformly to G_0 on the compact set Λ this gives a contradiction as before. □

The set of feasible systems does in general not depend in a fully continuous way on the observed spectrum. Therefore we use the weaker concept of upper semicontinuity. We call the set of feasible systems $\mathbf{B}(\Sigma; \delta, m, n)$ upper semicontinuous in (Σ, δ) if for all $(\Sigma_k, \delta_k) \to (\Sigma, \delta)$ and for $\mathcal{B}_k \in \mathbf{B}(\Sigma_k; \delta_k, m, n)$ with $\mathcal{B}_k \to \mathcal{B}_0$ there holds that $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta, m, n)$. As the sets of feasible systems are in general not compact, upper semicontinuity is not equivalent to the condition that $\rho(\mathbf{B}_k, \mathbf{B}_0) = \sup_{\mathcal{B}_k \in \mathbf{B}_k} \inf_{\mathcal{B}_0 \in \mathbf{B}_0} \mathrm{d}(\mathcal{B}_k, \mathcal{B}_0) \to 0$, where $\mathbf{B}_k := \mathbf{B}(\Sigma_k; \delta_k, m, n)$ and $\mathbf{B}_0 := \mathbf{B}(\Sigma; \delta, m, n)$. The following continuity results for feasible systems are valid for all factor schemes, that is, for the mean squares and uniform fit and for the cases with and without orthogonality constraint. We use the notation $\overline{\mathbf{B}}(\Sigma; \delta, m, n)$ for the set of all feasible systems for (Σ, δ) with m inputs and at most n states.

Proposition 14

(i) The set $\overline{\mathbf{B}}(\Sigma; \delta, m, n)$ is upper semicontinuous in (Σ, δ).

(ii) If n is minimal for (Σ, δ, m) then $\mathbf{B}(\Sigma; \delta, m, n)$ is upper semicontinuous in (Σ, δ).

(iii) Let n be minimal for $(\Sigma;\delta+\eta,m,n)$ for some $\eta>0$ and let $\mathbf{B}(\Sigma;\delta,m,n)$ be non-empty, then $\rho(\mathbf{B}(\Sigma_k;\delta_k,m,n),\mathbf{B}(\Sigma;\delta,m,n))\to 0$ if $(\Sigma_k,\delta_k)\to(\Sigma,\delta)$.

(iv) Under the conditions in (iii), $\mathbf{B}(\Sigma;\delta,m,n)$ is continuous from the right in δ.

Proof. (i) Let $(\Sigma_k,\delta_k)\to(\Sigma,\delta)$ and $\mathcal{B}_k\in\overline{\mathbf{B}}(\Sigma_k;\delta_k,m,n)$ with $\mathcal{B}_k\to\mathcal{B}$, then we have to prove that $\mathcal{B}\in\overline{\mathbf{B}}(\Sigma;\delta,m,n)$. That $\mathcal{B}$ has m inputs and at most n states follows from the fact that $\overline{\mathbf{B}}(m,n)$ is closed, see Proposition 11(ii). Further, Proposition 22 in the appendix implies that $\mathrm{d}(\Sigma_k,\mathcal{B}_k)\to\mathrm{d}(\Sigma,\mathcal{B})$, and this implies that $\mathrm{d}(\Sigma,\mathcal{B})\le\delta$ so that $\mathcal{B}\in\overline{\mathbf{B}}(\Sigma;\delta,m,n)$.

(ii) This corresponds to the situation in (i), where now $\mathcal{B}_k$ all have complexity (m,n). If $\mathcal{B}_k\to\mathcal{B}$ then $\mathcal{B}\in\overline{\mathbf{B}}(m,n)$ and $\mathrm{d}(\Sigma,\mathcal{B})\le\delta$. As n is minimal for (Σ,δ,m) it follows that $\mathcal{B}\in\mathbf{B}(m,n)$, so that $\mathcal{B}\in\mathbf{B}(\Sigma;\delta,m,n)$.

(iii) In a first step we prove that n is minimal for $(\Sigma_k,\delta+\eta,m)$ for all k large enough. If this were not true then there exist $n'<n$ and infinitely many indices k so that $\mathbf{B}(\Sigma_k;\delta+\eta,m,n')$ is not empty. For such indices let $\mathcal{B}_k\in\mathbf{B}(\Sigma_k;\delta+\eta,m,n')$ have minimal isometric representation (A_k,B_k,C_k,D_k), then the isometry condition implies that this sequence has a limit point, denoted by (A_0,B_0,C_0,D_0). Let $\mathcal{B}_0$ be the behaviour corresponding to these parameters, then $\mathcal{B}_0\in\mathbf{B}(m,n'')$ with $n''\le n'$. As in the proof of Proposition 13, the isometric kernel representations $\tilde{G}_k$ converge pointwise on the unit circle to the kernel representation $\tilde{G}_0$ of $\mathcal{B}_0$, except for a finite number of points. This implies that $\mathrm{d}(\Sigma_0,\mathcal{B}_0)\le\delta+\eta$, which contradicts the minimality of n for $(\Sigma,\delta+\eta,m)$. So n is minimal for $(\Sigma_k,\delta+\eta,m)$ and therefore $\mathbf{B}(\Sigma_k;\delta_k,m,n)$ is compact for k sufficiently large.

Now suppose that there exists an $\varepsilon>0$ and a sequence of systems $\mathcal{B}_k\in\mathbf{B}(\Sigma_k;\delta_k,m,n)$ so that $\mathrm{d}(\mathcal{B}_k,\mathcal{B})\ge\varepsilon$ for all $\mathcal{B}\in\mathbf{B}(\Sigma;\delta,m,n)$. As $\mathrm{d}(\Sigma_k,\mathcal{B}_k)\le\delta_k$ and $(\Sigma_k,\delta_k)\to(\Sigma,\delta)$ it follows from Proposition 22 in the appendix that for k sufficiently large $\mathcal{B}_k\in\mathbf{B}(\Sigma;\delta+\eta,m,n)$. As n is minimal for $(\Sigma,\delta+\eta,m)$ this is according to Proposition 13 a compact set, so the sequence $\mathcal{B}_k$ contains a limit point, say $\mathcal{B}_0\in\mathbf{B}(\Sigma;\delta+\eta,m,n)$. It follows from Proposition 22 that $\mathrm{d}(\Sigma,\mathcal{B}_0)\le\delta$ and thus $\mathcal{B}_0\in\mathbf{B}(\Sigma;\delta,m,n)$. From the assumption that $\mathrm{d}(\mathcal{B}_k,\mathcal{B})\ge\varepsilon$ for all $\mathcal{B}\in\mathbf{B}(\Sigma;\delta,m,n)$ this implies that $\mathrm{d}(\mathcal{B}_k,\mathcal{B}_0)\ge\varepsilon$, but this contradicts the fact that $\mathcal{B}_0$ is a limit point of the sequence $\mathcal{B}_k$.

(iv) Let $\delta_k\downarrow\delta$, then according to Proposition 13 the sets $\mathbf{B}(\Sigma;\delta_k,m,n)$ are compact for k sufficiently large. It follows from the result in (iii) that there holds $\rho(\mathbf{B}(\Sigma;\delta_k,m,n),\mathbf{B}(\Sigma;\delta,m,n))\to 0$, and as $\mathbf{B}(\Sigma;\delta,m,n)\subseteq\mathbf{B}(\Sigma;\delta_k,m,n)$ it is trivial that $\rho(\mathbf{B}(\Sigma;\delta,m,n),\mathbf{B}(\Sigma;\delta_k,m,n))=0$. This proves convergence in the Hausdorff metric. □

It is also of interest to consider the continuity of Pareto optimal models. Continuity in this respect is connected with robustness, in the sense that small perturbations in the data should lead to a small perturbation of optimal models.

We analyse this for models that optimize the fit under a complexity constraint. For given spectrum Σ we denote by $\mathbf{B}^*(\Sigma; m, n)$ the set of behaviours of optimally fitting factor models with m inputs and n states, and by $\overline{\mathbf{B}}^*(\Sigma; m, n)$ the set of optimally fitting behaviours with m inputs and at most n states.

Proposition 15

(i) The set $\overline{\mathbf{B}}^*(\Sigma; m, n)$ *is upper semicontinuous in the spectrum* Σ.

(ii) Let δ^* *be the optimal fit in* $\mathbf{B}(m, n)$ *and let* n *be minimal for* $(\Sigma, \delta^* + \eta, m)$ *for some* $\eta > 0$*, then* $\rho(\mathbf{B}^*(\Sigma_k; m, n), \mathbf{B}^*(\Sigma; m, n)) \to 0$ *for* $\Sigma_k \to \Sigma$.

Proof. (i) Let $\Sigma_k \to \Sigma$ and let $\mathcal{B}_k$ be an optimal behaviour in $\overline{\mathbf{B}}(m, n)$ for Σ_k with $\mathcal{B}_k \to \mathcal{B}$ for $k \to \infty$, then we have to prove that $\mathcal{B}$ is optimal for Σ. As $\overline{\mathbf{B}}(m, n)$ is closed it follows that $\mathcal{B} \in \overline{\mathbf{B}}(m, n)$, and if this limit system is not optimal then there exists a system $\mathcal{B}_0 \in \overline{\mathbf{B}}(m, n)$ so that $\mathrm{d}(\Sigma, \mathcal{B}_0) < \mathrm{d}(\Sigma, \mathcal{B})$. It then follows from Proposition 22 in the appendix that for k sufficiently large also $\mathrm{d}(\Sigma_k, \mathcal{B}_0) < \mathrm{d}(\Sigma_k, \mathcal{B}_k)$, but this contradicts the optimality of $\mathcal{B}_k$.

(ii) If this were not true then there exists an $\varepsilon > 0$ and a sequence of systems $\mathcal{B}_k \in \mathbf{B}^*(\Sigma_k; m, n)$ so that for all $\mathcal{B} \in \mathbf{B}^*(\Sigma; m, n)$ there holds $\mathrm{d}(\mathcal{B}_k, \mathcal{B}) \geq \varepsilon$. Now let $\mathcal{B} \in \mathbf{B}^*(\Sigma; m, n)$, so that $\mathrm{d}(\Sigma, \mathcal{B}) = \delta^*$ and $\mathrm{d}(\Sigma_k, \mathcal{B}) \leq \delta^* + \eta_k$ with $\eta_k \downarrow 0$ for $k \to \infty$. It then follows that $\mathrm{d}(\Sigma_k, \mathcal{B}_k) \leq \delta^* + \eta_k$ and hence $\mathrm{d}(\Sigma, \mathcal{B}_k) \leq \delta^* + \eta$ for k sufficiently large. Because n is minimal for $(\Sigma, \delta^* + \eta, m)$ it follows that $\mathbf{B}(\Sigma; \delta^* + \eta, m, n)$ is compact, so that the sequence $\mathcal{B}_k$ has a limit point, say $\mathcal{B}_0 \in \mathbf{B}(m, n)$. As $\mathrm{d}(\mathcal{B}_k, \mathcal{B}) \geq \varepsilon$ for all $\mathcal{B} \in \mathbf{B}^*(\Sigma; m, n)$ the same holds true for $\mathcal{B}_0$, but this contradicts the fact that $\mathrm{d}(\Sigma, \mathcal{B}_0) = \lim \mathrm{d}(\Sigma_k, \mathcal{B}_k) = \delta^*$ so that $\mathcal{B}_0 \in \mathbf{B}^*(\Sigma; m, n)$. □

4.3 Consistency

Next we investigate the consistency of dynamic factor models when the spectrum is estimated from observed data. In applications the spectrum of the observed process will in general be unknown. Suppose that, apart from assumptions A1-A4, the available information on the process consists of an observed time series of length T. Let Σ_T denote an estimator of the process spectrum Σ that is based on this time series. In order to simplify the analysis we assume that the estimator is strongly consistent, so that $\mathrm{d}(\Sigma, \Sigma_T) \to 0$ almost surely for $T \to \infty$. A strongly consistent estimator can be obtained, for example, as follows. Let the observed process have spectrum $\Sigma(z) = \sum_{k=-\infty}^{\infty} R(k) z^{-k}$ where $R(k) := \mathrm{E}\{w(t)w'(t-k)\}$ are the process covariances, and let $\hat{R}_T(k) = \frac{1}{T} \sum_{t=k+1}^{T} w(t)w'(t-k)$ be the sample covariances.

Proposition 16 *Under weak conditions on the data generating process, a strongly consistent estimator of* Σ *is given by* $\Sigma_T(z) = \sum_{|k| \leq k_T} \hat{R}_T(k) z^{-k}$, *where* $k_T = log(T)$.

Proof. The estimation error is bounded by

$$\| \Sigma(z) - \Sigma_T(z)\|_\infty \leq (2k_T + 1) \sup_{|k|\leq k_T} \|R(k) - \hat{R}_T(k)\| + \sum_{|k|>k_T} \|R(k)\|.$$

The second term converges to zero by Assumption A4, and the first term converges to zero almost surely under weak conditions. A sufficient condition is that the spectrum Σ is rational, but the result holds also true for a broad class of nonrational spectra. For these results we refer to [13, Theorems 5.3.2 and 7.4.3]. □

In the following let $\mathbf{B}_0 := \mathbf{B}(\Sigma; \delta, m, n)$ be the class of feasible models and $\mathbf{B}_0^* \subset \mathbf{B}(m,n)$ the set of optimal models of complexity (m,n), that is, with optimal fit in this class. By $\overline{\mathbf{B}}_0$ and $\overline{\mathbf{B}}_0^*$ we denote the sets of feasible and optimal models respectively with m inputs and at most n states. Further let $\mathbf{B}_T := \mathbf{B}(\Sigma_T; \delta, m, n)$ be the set of feasible models and $\mathbf{B}_T^*$ the set of optimal models of complexity (m,n) for the estimated spectrum Σ_T, and let $\overline{\mathbf{B}}_T$ and $\overline{\mathbf{B}}_T^*$ be the sets of feasible and optimal models respectively with m inputs and at most n states. These are random sets as they depend on the observed time series. The next two theorems state consistency properties for feasible and optimal models, where it is assumed that the estimator Σ_T is strongly consistent.

Theorem 17

(i) *Behaviours with better fit than the noise bound are estimated consistently, that is, if a factor model has behaviour $\mathcal{B}$ of complexity (m,n) and fit δ then for $\delta' > \delta$ there holds almost surely that $\mathcal{B} \in \mathbf{B}(\Sigma_T; \delta', m, n)$ for $T \to \infty$.*

(ii) *The sample estimator of the set of feasible behaviours in $\overline{\mathbf{B}}(m,n)$ is upper semiconsistent, in the sense that $\{\mathcal{B}_T \in \overline{\mathbf{B}}_T, \mathcal{B}_T \to \mathcal{B}_0\} \Rightarrow \{\mathcal{B}_0 \in \overline{\mathbf{B}}_0\}$ holds almost surely, that is, the set of data with this convergence property has probability one.*

(iii) *If n is minimal for $(\Sigma, \delta+\eta, m)$ for some $\eta > 0$, then the set of feasible sample behaviours in $\mathbf{B}(m,n)$ converges to a subset of the feasible behaviours for the process, in the sense that $\rho(\mathbf{B}_T, \mathbf{B}_0) \to 0$ almost surely for $T \to \infty$.*

Proof. (i) This evident as $\Sigma_T \to \Sigma$ almost surely and $\mathrm{d}(\Sigma, \mathcal{B})$ is continuous, see Proposition 22 in the appendix.

(ii) This follows from Proposition 14(i).

(iii) This follows from Proposition 14(iii). □

Theorem 18

(i) *The sample estimator of the set of optimal behaviours in* $\overline{\mathbf{B}}(m,n)$ *is upper semiconsistent, in the sense that* $\{\mathcal{B}_T \in \overline{\mathbf{B}}^*_T, \mathcal{B}_T \to \mathcal{B}_0\} \Rightarrow \{\mathcal{B}_0 \in \overline{\mathbf{B}}^*_0\}$ *almost surely.*

(ii) *If the process spectrum has a unique optimal factor behaviour* $\mathcal{B}^*_0$ *of complexity* (m,n) *and if the infimum of the fits of models in* $\overline{\mathbf{B}}(m,n-1)$ *is strictly larger than the fit of* $\mathcal{B}^*_0$*, then this behaviour is estimated consistently in the sense that* $\mathrm{d_H}(\mathbf{B}^*_T, \{\mathcal{B}^*_0\}) \to 0$ *almost surely for* $T \to \infty$*.*

Proof. (i) This follows from Proposition 15(i).

(ii) As it is given that $\mathbf{B}^*_0 = \{\mathcal{B}^*_0\}$ is a singleton it follows that $\rho(\{\mathcal{B}^*_0\}, \mathbf{B}^*_T) = \inf_{\mathcal{B}\in\mathbf{B}^*_T} \mathrm{d}(\mathcal{B}^*_0, \mathcal{B}) \leq \sup_{\mathcal{B}\in\mathbf{B}^*_T} \mathrm{d}(\mathcal{B}^*_0, \mathcal{B}) = \rho(\mathbf{B}^*_T, \{\mathcal{B}^*_0\})$, so it suffices to prove that the last expression converges to zero. Let the optimal fit for Σ_T among models of complexity (m,n) be given by δ^*_T and let $\delta^*_0 = \mathrm{d}(\Sigma, \mathcal{B}^*_0)$, then it follows from $\mathrm{d}(\Sigma_T, \mathcal{B}^*_0) \to \delta^*_0$ that $\delta^*_T \to \delta^*_0$ almost surely. Further, because of the assumption that $\inf\{\mathrm{d}(\Sigma, \mathcal{B}); \mathcal{B} \in \overline{\mathbf{B}}(m, n-1)\} > \delta^*_0$, it follows that n is minimal for all $(\Sigma, \delta^*_0 + \eta, m)$ with $\eta \geq 0$ sufficiently small, and the same holds then true almost surely for $(\Sigma_T, \delta^*_T + \eta, m)$ if $T \to \infty$. Then for T sufficiently large $\mathbf{B}^*_T$ is a closed subset of the compact set $\mathbf{B}(\Sigma_T; \delta^*_T + \eta, m, n)$, so that $\mathbf{B}^*_T$ is compact. This means that the Hausdorff distance is well-defined. Further, as $(\Sigma_T, \delta^*_T) \to (\Sigma, \delta^*_0)$ almost surely it follows from Proposition 14(iii) that

$$\rho(\mathbf{B}^*_T, \{\mathcal{B}^*_0\}) = \rho(\mathbf{B}(\Sigma_T; \delta^*_T, m, n), \mathbf{B}(\Sigma; \delta^*_0, m, n)) \to 0 \text{ almost surely.}$$

□

This means that, under the above conditions, the feasible and optimal finite sample models are in the limit also feasible and optimal for the data generating process. However, possibly not all feasible and optimal models are identified in this way.

4.4 Low Noise Consistency

We conclude our analysis by considering another kind of consistency, inspired by the concept of low noise as defined in [15]. This is based on the idea that an identification method which aspires to deal with noisy data must, as a minimal requirement, function well when dealing with data having low noise content. Let the observed process be given by $w = \hat{w}_0 + \tilde{w}_0$, where the latent process $\hat{w}_0$ is fixed and has behaviour $\mathcal{B}_0$ of complexity (m_0, n_0) and where the noise process $\tilde{w}_0$ has norm δ_0. Low noise consistency corresponds to the condition that the factor behaviour $\mathcal{B}_0$ is identified uniquely if the noise vanishes in the limit. The following result shows that this holds true, provided that the factor scheme is specified correctly.

Proposition 19

- (i) *If the factor scheme, noise bound and complexity have been specified correctly then the factor behaviour is identified, that is, if $\delta \geq \delta_0$, $m = m_0$ and $n = n_0$ then $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta, m, n)$. If orthogonality is imposed but the data generating process does not satisfy this property, then the system need not be identified.*

- (ii) *Correctly specified factor schemes are low noise consistent, that is, if $\delta_0 \leq \delta \downarrow 0$ then the set of feasible behaviours $\mathbf{B}(\Sigma; \delta, m, n) \to \{\mathcal{B}_0\}$ (in the sense of the Hausdorff metric) for $(m, n) = (m_0, n_0)$, and $\mathbf{B}(\Sigma; \delta, m, n) \to \emptyset$ if $m < m_0$ or $m = m_0, n < n_0$. Consistency is in general lost if orthogonality is imposed but the data generating process does not satisfy this property.*

Proof. (i) This is evident from the definition of $\mathbf{B}(\Sigma; \delta, m, n)$.

(ii) The process decomposition $w = \hat{w}_0 + \tilde{w}_0$ induces a corresponding spectral decomposition $\Sigma = \hat{\Sigma}_0 + \tilde{\Sigma}_0 + \Sigma_c + {\Sigma_c}'$ where $\hat{\Sigma}_0$ is the spectrum of the latent process $\hat{w}_0$, $\tilde{\Sigma}_0$ of the noise $\tilde{w}_0$, and Σ_c is the cross spectrum between $\hat{w}_0$ and $\tilde{w}_0$. As the latent process $\hat{w}_0$ is fixed and the noise converges to zero, it follows that $\| \Sigma - \hat{\Sigma}_0 \|_\infty = \| \tilde{\Sigma}_0 + \Sigma_c + {\Sigma_c}' \|_\infty \to 0$.

First we consider the factor scheme without orthogonality constraint. Then the misfit function $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B})$ is also well-defined for the singular spectral density $\hat{\Sigma}_0$, i.e., if P is the projection onto $\mathcal{B}$ and $\tilde{\Sigma} = (I - P)\,\hat{\Sigma}_0(I - P)$ then $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}) = \| \tilde{\Sigma}^{1/2} \|$ and $\mathbf{B}(\hat{\Sigma}_0; \delta, m, n) = \{\mathcal{B} \in \mathbf{B}(m, n) |\, \mathrm{d}(\hat{\Sigma}_0, \mathcal{B}) \leq \delta\}$. It can easily be shown, along the lines of the proof of Proposition 22 in the appendix, that $\mathrm{d}(\Sigma, \mathcal{B}) \to \mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_*)$ if $\Sigma \to \hat{\Sigma}_0$ and $\mathcal{B} \to \mathcal{B}_*$. In addition there holds

$$\left|\mathrm{d}^2(\Sigma, \mathcal{B}) - \mathrm{d}^2(\hat{\Sigma}_0, \mathcal{B})\right| \leq c \| \Sigma - \hat{\Sigma}_0 \|_\infty$$

where $c = 2$ for the uniform norm and $c = 2\pi q$ for the mean squares norm. The above result follows from the proof of Lemma 21 in the appendix and the inequality $\|(I - P)(\Sigma - \hat{\Sigma}_0)(I - P)\|_\infty \leq \| \Sigma - \hat{\Sigma}_0 \|_\infty$.

We now first show that for $m < m_0$ or $m = m_0$, $n < n_0$, the infimum of the misfits $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B})$ over the set of behaviours $\mathbf{B}(m, n)$ is strictly larger than zero. If this were not true, then there would exist a sequence of behaviours $\mathcal{B}_k \in \mathbf{B}(m, n)$, with corresponding projections P_k, such that $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_k) \to 0$. As in the proof of Proposition 13, it follows that there exists a subsequence $k(l)$ and a behaviour $\mathcal{B}_* \in \mathbf{B}(m, n'), n' \leq n$, with a corresponding projection P_*, such that $P_{k(l)}(z) \to P_*(z)$ for $l \to \infty$, pointwise on the unit circle except for a finite number of points. Then $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_k) \to 0$ implies that $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_*) = 0$, and this means that $\mathcal{B}_0 \subseteq \mathcal{B}_*$. This contradicts the assumption that the complexity (m, n) is smaller than the complexity (m_0, n_0) of $\mathcal{B}_0$. We conclude that the infimum of misfits of models of complexity $m < m_0$ or $m = m_0$, $n < n_0$ is given by a strictly positive number δ_*. Since $\| \Sigma - \hat{\Sigma}_0 \|_\infty$ converges to zero for $\delta \downarrow 0$, there exists a $\delta_+ > 0$ such that $c \| \Sigma - \hat{\Sigma}_0 \|_\infty < \delta_*^2$ for $\delta \leq \delta_+$. By the above considerations and inequalities, there holds for $\delta \leq \delta_+$ that

$$\mathrm{d}^2(\Sigma, \mathcal{B}) \geq \mathrm{d}^2(\hat{\Sigma}_0, \mathcal{B}) - c \| \Sigma - \hat{\Sigma}_0 \|_\infty > \delta_*^2 - \delta_*^2 = 0.$$

This shows that $\mathbf{B}(\Sigma;\delta,m,n)$ is empty for $m < m_0$ and for $m = m_0$, $n < n_0$ if $\delta \leq \delta_+$.

Now suppose that the complexity has been specified correctly. In this case $\mathcal{B}_0 \in \mathbf{B}(\Sigma;\delta,m_0,n_0)$ so that $\rho(\{\mathcal{B}_0\},\mathbf{B}(\Sigma;\delta,m_0,n_0)) = 0$. Further, from the foregoing it follows that n_0 is minimal for $(\hat{\Sigma}_0;\delta_+,m_0)$, as $\mathbf{B}(\hat{\Sigma}_0;\delta_+,m_0,n) = \emptyset$ for $n < n_0$ and $\mathbf{B}(\hat{\Sigma}_0;\delta_+,m_0,n_0)$ is not empty, and also $\mathbf{B}(\hat{\Sigma}_0;0,m_0,n_0) = \{\mathcal{B}_0\}$. It follows from Proposition 14(iii) that $\rho(\mathbf{B}(\Sigma;\delta,m_0,n_0),\{\mathcal{B}_0\}) \to 0$.

Next we consider the factor scheme with orthogonality. By imposing the orthogonality constraint the sets $\mathbf{B}(\Sigma;\delta,m,n)$ in general become smaller. Since $\mathcal{B}_0 \in \mathbf{B}(\Sigma;\delta,m_0,n_0)$ for $\delta_0 \leq \delta$, the above results imply that $\mathbf{B}(\Sigma;\delta,m,n) \to \emptyset$ if the complexity (m,n) is smaller than (m_0,n_0) and that $\mathbf{B}(\Sigma;\delta,m_0,n_0) \to \{\mathcal{B}_0\}$.

That consistency is lost if orthogonality is imposed but the data generating process is not orthogonal is evident from Theorem 5(i), as this shows that in this case the misfit δ_0 can in general not be obtained in the class of orthogonal models in $\mathbf{B}(m,n)$. □

4.5 Illustration

We will illustrate the foregoing results for static factor models, as in this case more explicit characterizations can be obtained. We will not further discuss the dynamic system example of Sections 2.4 and 3.3, as the consistency analysis for dynamic factor models will be the topic of another paper.

So assume that the observed proces w is white noise, and let Σ denote the covariance matrix of w. As we have seen in Section 3.3.1, we can without loss of fit restrict ourselves to static relations. The set of all static systems $\mathbf{B}(m,0)$ is isomorphic to the set of all m-dimensional linear subspaces of $\mathbf{R}^q$. Isometric kernel representations of static systems are isometric matrices $\tilde{G} \in \mathbf{R}^{q\times m}$.

It can easily be seen that $\mathcal{B} \in \mathbf{B}(\Sigma;\delta,m,0)$ if and only if the isometry $\tilde{G}$ satisfies the following inequalities: for the non-orthogonal factor scheme, $\mathrm{trace}(\tilde{G}'\,\Sigma\,\tilde{G}) \leq \delta^2$ for the mean squares norm and $\tilde{G}'\,(\Sigma - \delta^2 I)\,\tilde{G} \leq 0$ for the uniform norm, and for the orthogonal factor scheme $\mathrm{trace}(\Sigma\,\tilde{G}(\tilde{G}'\,\Sigma\,\tilde{G})^{-1}\,\tilde{G}'\,\Sigma) \leq \delta^2$ and $\tilde{G}'(\Sigma^2 - \delta^2\,\Sigma)\,\tilde{G} \leq 0$ respectively. From this characterization it follows that the sets $\mathbf{B}(\Sigma;\delta,m,0)$ of static systems are always compact.

Let $\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0$ denote the eigenvalues of Σ, and let $\Sigma = \hat{\Sigma}_m + \tilde{\Sigma}_m$ be the principal component decomposition of Σ with m factors as in (15). The set $\mathbf{B}(\Sigma;\delta,m,0)$ is nonempty if and only if $\|\,\tilde{\Sigma}_m\,\| \leq \delta$, that is, $\lambda_{m+1}^{1/2} \leq \delta$ for the uniform norm and $(\lambda_{m+1} + \ldots + \lambda_q)^{1/2} \leq \delta$ for the mean squares norm. Furthermore one can show that the sets $\mathbf{B}(\Sigma;\delta,m,0)$ depend continuously on (Σ,δ) with the exception of points where $\|\,\tilde{\Sigma}_m\,\| = \delta$.

Let Σ_T denote a strongly consistent estimator of Σ. If $\|\,\tilde{\Sigma}_m\,\| < \delta$ then $\mathbf{B}(\Sigma_T;\delta,m,0)$ is a strongly consistent estimator of $\mathbf{B}(\Sigma;\delta,m,0)$. The principal component model of Σ_T is a strongly consistent estimator of the principal component model of Σ, so that the Pareto optimal models are estimated consistently.

5 Conclusion

Dynamic factor models decompose an observed process in terms of an underlying latent component and additional noise. The variables are treated in a completely symmetric way, and no assumptions on inputs and outputs are required. The latent process has a singular spectrum as it satisfies deterministic dynamic relationships. This means that the factor behaviour consists of a linear dynamical system. In particular, the latent process has less free variables than the observed process. Depending on the chosen factor scheme, several interpretations of the noise process are possible. If the noise can be assumed to be uncorrelated with the factor process this is called the orthogonal factor scheme. This is the usual assumption in the classical models of factor analysis. In other situations it is more natural to consider the latent process as an approximation of the observed process and to assume that the factor components are constructed from the observations. This is called the observable factor scheme.

Within this framework we investigated the representation of dynamic factor models and defined notions of complexity and goodness of fit. Concerning the identification of factor models we presented characterizations of Pareto optimal models and we derived results on consistency, both in case of observed data and in case of low noise.

An advantage of our approach is that it deals explicitly with the symmetric modelling of observed data by means of dynamic stochastic models. Other contributions in symmetric system modelling have been developed in the behavioural identification of systems and in the structural analysis of factor models. In a sense, our approach can be seen as an extension of these two frameworks. It enriches the deterministic behavioural framework with a stochastic analysis, and it extends the traditionally structure oriented analysis of factor models to a more empirical modelling setting.

Several questions deserve further investigation. Of special interest is the analysis of identification procedures within this framework. Another issue is the incorporation of prior knowledge, for example concerning the input-output structure of the model. A further analysis of the probabilistic structure of factor models is needed in order to develop statistical test procedures, for example to estimate the complexity of factor models from observed data.

6 Appendix

Lemma 20 *Let $A, B \in \mathbf{C}^{q\times q}$ be two positive semidefinite matrices and let $\lambda_1(A) \geq \cdots \geq \lambda_q(A) \geq 0$ and $\lambda_1(B) \geq \cdots \geq \lambda_q(B) \geq 0$ be the eigenvalues of A and B respectively. Then*

(i) $|\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|_\infty$

(ii) For every unitary matrix $U \in \mathbf{C}^{q\times m}$, $U^*U = I$, there holds

$$\begin{array}{rcl} \mathrm{trace}(UU^*AUU^*) & = & \mathrm{trace}(UAU^*) \geq \lambda_{m+1}(A) + \cdots \lambda_q(A) \\ \lambda_{\max}(UU^*AUU^*) & = & \lambda_{\max}(UAU^*) \geq \lambda_{m+1}(A) \end{array}$$

The lower bound is reached if the columns of U form a basis for the eigenspace of A corresponding to the $q - m$ smallest eigenvalues.

Proof. See [12, Corollary 8.1.3 and Theorem 8.1.2]. □

Lemma 21 *Let Σ_k be a sequence of spectral densities that converges to Σ_0 in the sense that $\| \Sigma_k - \Sigma_0 \|_\infty \to 0$. Then*

(i) $\| \Sigma_k^{1/2} \| \to \| \Sigma_0^{1/2} \|$.

(ii) If Σ_0 is positive definite, then Σ_k is positive definite for all k sufficiently large and $\| \Sigma_k^{-1} - \Sigma_0^{-1} \|_\infty \to 0$.

Proof. (i) By Lemma 20 $|\lambda_i(\Sigma_k(z)) - \lambda_i(\Sigma_0(z))| \leq \| \Sigma_k - \Sigma_0 \|_\infty$ pointwise on the unit circle, so that

$$\begin{array}{rcl} |\| \Sigma_k^{1/2} \|_2^2 - \| \Sigma_0^{1/2} \|_2^2| & = & | \oint_{|z|=1} \mathrm{trace}(\Sigma_k(z) - \Sigma_0(z)) dz| \\ & \leq & 2\pi q \| \Sigma_k - \Sigma_0 \|_\infty \\ |\| \Sigma_k^{1/2} \|_\infty^2 - \| \Sigma_0^{1/2} \|_\infty^2| & = & | \sup_{|z|=1} \lambda_{\max}(\Sigma_k(z)) - \sup_{|z|=1} \lambda_{\max}(\Sigma_0(z))| \\ & \leq & 2\| \Sigma_k - \Sigma_0 \|_\infty \end{array}$$

(ii) By the assumption $\Sigma_0 > 0$ and the result in Lemma 20 for the eigenvalues of Σ_k, it follows that $\| \Sigma_0^{-1} \|_\infty = 1/\{\inf_{|z|=1} \lambda_{\min}(\Sigma_0(z))\}$ and $\| \Sigma_k^{-1} \|_\infty$ are bounded. The result then follows from

$$\| \Sigma_k^{-1} - \Sigma_0^{-1} \|_\infty = \| \Sigma_k^{-1}(\Sigma_0 - \Sigma_k) \Sigma_0^{-1} \|_\infty \leq \| \Sigma_k^{-1} \|_\infty \|(\Sigma_0 - \Sigma_k)\|_\infty \| \Sigma_0^{-1} \|_\infty$$

□

Proposition 22 *The misfit function* $\mathrm{d}(\Sigma, \mathcal{B})$ *is continuous in* $(\Sigma, \mathcal{B})$ *for all positive definite spectral densities* Σ.

Proof. Let $\Sigma_k \to \Sigma_0 > 0$ and $\mathcal{B}_k \to \mathcal{B}_0$ be convergent sequences of spectral densities and behaviours respectively. The corresponding isometric kernel representations of $\mathcal{B}_k$, $\mathcal{B}_0$ are denoted by $\tilde{G}_k$ and $\tilde{G}_0$ respectively. The optimal noise spectra, given in Theorem 5, corresponding to the spectral densities Σ_k, Σ_0 and the behaviours $\mathcal{B}_k$, $\mathcal{B}_0$ are denoted by $\tilde{\Sigma}_k$ and $\tilde{\Sigma}_0$ respectively. By Lemma 21 it suffices to show that $\| \tilde{\Sigma}_k - \tilde{\Sigma}_0 \|_\infty \to 0$.

For the case without orthogonality the noise spectra are given by $\tilde{\Sigma}_k = \tilde{G}_k \, \tilde{G}_k^* \, \Sigma_k \, \tilde{G}_k \, \tilde{G}_k^*$ and $\tilde{\Sigma}_0 = \tilde{G}_0 \, \tilde{G}_0^* \, \Sigma_0 \, \tilde{G}_0 \, \tilde{G}_0$, in which case $\| \tilde{\Sigma}_k - \tilde{\Sigma}_0 \|_\infty \to 0$ is evident.

For the case with orthogonality, let $\bar{G}_k = \tilde{G}_k \, \tilde{G}_k^* \, \tilde{G}_0$, then $\|\bar{G}_k - \tilde{G}_0 \|_\infty \leq \| \tilde{G}_k \, \tilde{G}_k^* - \tilde{G}_0 \, \tilde{G}_0^* \|_\infty \| \tilde{G}_0 \|_\infty \to 0$. The noise spectra for this factor scheme are given by $\tilde{\Sigma}_0 = \Sigma_0 \, \tilde{G}_0 (\tilde{G}_0^* \, \Sigma_0 \, \tilde{G}_0)^{-1} \, \tilde{G}_0^* \, \Sigma_0$ and $\tilde{\Sigma}_k = \Sigma_k \, \tilde{G}_k (\tilde{G}_k^* \, \Sigma_k \, \tilde{G}_k)^{-1} \, \tilde{G}_k^* \, \Sigma_k = \Sigma_k \, \bar{G}_k (\bar{G}_k^* \, \Sigma_k \, \bar{G}_k)^{-1} \bar{G}_k^* \, \Sigma_k$, where the last equality follows form the fact that $\tilde{G}_k^* \, \tilde{G}_0 \to I$ so that this is invertible for k sufficiently large. The result now follows from Lemma 21. $\square$

References

[1] T.W. Anderson and H. Rubin, Statistical inference in factor analysis, in J. Neyman (ed.), *Proceedings Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1956, pp. 111-150.

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

[3] D.R. Brillinger, *Time Series, Data Analysis and Theory*, Holden-Day, 1981.

[4] P.E. Caines, *Linear Stochastic Systems*, Wiley, 1988.

[5] M. Deistler, Symmetric modeling in system identification, in H. Nijmeijer and J.M. Schumacher (eds.), *Three Decades of Mathematical System Theory*, Springer, 1989, pp. 128-147.

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

[7] M. Deistler and W. Scherrer, System identification and errors in the variables, in K.Haagen et al. (eds.), *Statistical Modelling and Latent Variables*, North-Holland, 1993, pp. 95-111.

[8] R.F. Engle and M. Watson, A one-factor multivariate time series model of metropolitan wage rates, *Journal of the American Statistical Association* 76, 1981, pp. 774-781.

[9] R. Frisch, *Statistical Confluence Analysis by means of Complete Regression Systems*, Publ. 5, Economic Institute, University of Oslo, 1934.

[10] W.A. Fuller, *Measurement Error Models*, Wiley, 1987.

[11] J.F. Geweke, The dynamic factor analysis of economic time series models, in D.J. Aigner and A.S. Goldberger (eds.), *Latent Variables in Socio-economic Models*, North Holland, 1977, pp. 365-383.

[12] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1989.

[13] E.J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*, Wiley, 1988.

[14] C. Heij, *Deterministic Identification of Dynamical Systems*, Springer, 1989.

[15] R.E. Kalman, A theory for the identification of linear relations, in H. Brezis et al. (eds.), *Colloques Lions*, 1989.

[16] E.L. Leamer, Errors in variables in linear systems, *Econometrica* 55, 1987, pp. 893-909.

[17] L. Ljung, *System Identification : Theory for the User*, Prentice-Hall, 1987.

[18] G. Picci and S. Pinzoni, Dynamic factor analysis models for stationary processes, *IMA Journal of Mathematical Control and Information* 3, 1986, pp. 185-210.

[19] M.B. Priestley, *Spectral Analysis and Time Series*, Academic Press, 1981.

[20] B. Roorda, Algorithms for global total least squares modelling of finite multivariable time series, *Report*, Tinbergen Institute, Erasmus University Rotterdam, 1994. To appear in Automatica.

[21] B. Roorda and C. Heij, Global total least squares modelling of multivariable time series, *Report* 9343, Econometric Institute, Erasmus University Rotterdam, 1993. To appear in IEEE Transactions on Automatic Control.

[22] Y.A. Rozanov, *Stationary Random Processes*, Holden Day, 1967.

[23] J.H. van Schuppen, Stochastic realization problems, in H. Nijmeijer and J.M. Schumacher (eds.), *Three Decades of Mathematical System Theory*, Springer, 1989, pp. 480-523.

[24] J.C. Willems, From time series to linear system, part I, *Automatica* 22, 1986, pp. 561-580.

[25] J.C. Willems, From time series to linear system, part III, *Automatica* 23, 1987, pp. 87-115.

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