

Consistency of Global Total Least Squares in Stochastic System Identification

C. Heij *

W. Scherrer †

Abstract

Global total least squares has been introduced as a method for the identification of deterministic system behaviours. We analyse this method within a stochastic framework, where the observed data are generated by a stationary stochastic process. Conditions are formulated so that the method is consistent in the sense that, when the number of observations tends to infinity, the identified deterministic behaviour converges to the behaviour that provides an optimal approximation of the data generating process.

Keywords: Linear systems, system behaviour, factor analysis, least squares, consistency.

1 Problem Statement

System identification is concerned with the determination of sufficiently simple models that give a sufficiently accurate description of the observed data. Identification methods differ in the specification of the model class and in the way the complexity and accuracy of models is evaluated. Within the area of systems and control, a well-known and much used approach is that of prediction error identification of linear input-output systems, see [9], or more general, the maximum likelihood estimation of ARMAX systems, see [3]. These methods require that several system properties have been specified prior to model estimation.

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

†Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Argentinierstrasse 8, A-1040 Wien, Austria, phone +43-1-58801-4405, fax +43-1-5054524, e-mail wo_sch@e119ws1.tuwien.ac.at.

tion. This concerns in particular the decomposition of the system variables into inputs and outputs, and the joint correlation structure between inputs, outputs and disturbances.

In practice it is not always clear which variables act as inputs and which ones as outputs, and often the properties of the model errors are unknown. Several approaches have been developed for a more symmetric modelling of linear systems, in particular errors-in-variables models, see [1], [2], and system behaviours, see [13], [14], [4], [10]. Errors-in-variables models treat the system variables in a symmetric way, but rather strong noise assumptions are needed. The behavioural approach is of a completely deterministic nature that prevents any asymmetry but also lacks a statistical analysis of identification procedures. A synthesis of the two approaches is proposed in [8] in terms of a new class of dynamic factor models, see also [5] and [6]. Hereby it is assumed that the observations are generated by a stationary stochastic process, denoted by w . A factor model is a decomposition

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

where θ satisfies deterministic linear dynamic restrictions. This is called the factor process, and ϵ is the corresponding error process. Factor models are evaluated in terms of their complexity, measured by the number of degrees of freedom of the factor process, and in terms of the magnitude of the error. This is made more specific in the next section. Fixing the allowed complexity, a model is called optimal if it has minimal error under this constraint.

In this paper we consider the question of consistency. That is, the question whether optimal models can be identified from observed data in the limit when the length of the observation interval tends to infinity. This is investigated for the so-called global total least squares identification method, see [12] and especially [10]. This method is described in Section 3, a

parametrization in Section 4, and consistency results in Section 5. In Section 2 we introduce the model class, further details of which are given in [8].

2 Factor Models

The traditional model of static factor analysis is of the form $w = Lf + v$, where w is a vector of observed variables, f a lower dimensional vector of unobserved latent variables, v a vector of unobserved noise components, and L a matrix of factor loadings. If we define the factor variables by $f = L^{-1}w$, then these variables satisfy deterministic linear equations as the matrix L does not have full row rank. In analogy with this, we define a dynamic factor model for the process w as in facmod, where the factor process satisfies deterministic linear difference equations. That is,

$$R(\sigma) = 0 \quad (2)$$

where σ denotes the shift operator and R is a polynomial matrix in the shift. The behaviour corresponding to these equations is defined as the set of all time series that satisfy the equations ar. This behaviour corresponds to a linear system, see [13], [14]. The behaviour of the factor model facmod is defined by the maximal set of equations ar, that is, it is the smallest linear system that contains almost all factor realizations in the sense that the set $\{\omega \in \Omega; (\omega) \in\}$ has probability one. We will restrict the attention to factor processes that are purely non-deterministic, in the sense of stochastic processes. In this case the behaviour is controllable and corresponds to an input-output system that can be described by a rational transfer function matrix. This is the usual model class in system identification, but in our approach the variables are not distinguished in inputs and outputs and the behaviour is defined in terms of system trajectories, not in terms of a parametric description such as a transfer function.

The complexity of a factor model facmod is defined in terms of its behaviour. Let this correspond to a system with m inputs and n states, then the complexity is defined as the pair (m, n) . If q denotes the total number of variables in the system, then a factor model of complexity (m, n) can be described by equations ar, where $(q - m)$ is the number of independent equations and n the (smallest) sum of the orders of each of the individual equations corresponding to the rows of R . A behaviour is less complex the smaller m and n are, that is, the less (unrestricted)

inputs and initial conditions it allows. Stated otherwise, simple factor models are restricted by many equations of short lag.

In the sequel let the observed process w be fixed. Imposing more restrictions on the factor process will in general result in a larger error. This is measured by the mean squared error, that is, $\{E\|t\|^2\}^{1/2}$ where $\|\cdot\|$ denotes the Euclidean norm in q -dimensional space. The error of a behaviour with respect to the process w is defined as the smallest achievable error by factor models with behaviour. The corresponding factor model is given by $f = Pw$ and $v = (I - P)w$, where P is the operator of orthogonal projection onto, see [8]. If the process w has spectrum ξ that is bounded on the unit circle, then the error is given by

$$e() = \left[\int_{|z|=1} \text{tr}\{(I - P(z))\xi(z)\} dz \right]^{1/2} \quad (3)$$

where tr denotes the trace of a matrix. For fixed complexity (m, n) , a behaviour is called optimal if it minimizes the error erbehav under this complexity constraint. The problem of determining optimal models for given spectrum ξ and complexity (m, n) has been solved in [8], see also [5]. This basically involves the l_2 -approximation of a spectral factor of ξ . For l_2 -approximation algorithms we refer to [11] and [12].

3 Global Total Least Squares

We now consider the situation where the process w is unknown, and the available information consists of an observed time series generated by the process. That is, the data $w_N = w(\omega)_{|[1, N]}$ consists of a realization $w(\omega)$ of the process observed on a time interval of length N . We define the error of a behaviour with respect to these data as the global total least squares (GTLS) distance. That is,

$$e_N() = \min \left\{ \frac{1}{\sqrt{N}} \|w_N - w_N^a\| \right\} \quad (4)$$

with $w_N^a \in_{|[1, N]}$ and with $\|\cdot\|$ the Euclidean norm in (qN) -dimensional space. This distance involves the total squares, in the sense that approximations in all the variables are allowed. It is also global in the sense that the approximation w_N^a should not only locally satisfy the equations ar, as in prediction oriented criteria, but also globally, as the full trajectory

w_N^a should satisfy the laws of the behaviour . For further motivation and for algorithms we refer to [12] and, in particular, to [10].

For fixed complexity (m, n) , a behaviour is called optimal for the observed data if it minimizes the error GTLS under this complexity constraint. It can be shown that, with the exclusion of exceptional cases, this optimal behaviour exists and is unique, and we denote it by $_N(m, n)$. This behaviour is a random set as it depends on the realization of the observed process, that is, on $\omega \in \Omega$. Let (m, n) be the optimal model in the sense of minimizing erbehav under the complexity constraint, that is, giving the minimal achievable error if the data generating process is known. The basic question of this paper is in which sense and under which conditions it holds true that, for almost all $\omega \in \Omega$,

$$\lim_{N \rightarrow \infty} {}_N(m, n) = (m, n). \quad (5)$$

In this case the global total least squares method is consistent.

4 Parametrization

The foregoing exposition was focused on behavioural system properties. The complexity and accuracy of behaviours is expressed in terms of set theoretic properties of the system. For computational purposes it is of course essential to use parametric representations, and these can now be chosen as it suits.

For the GTLS error GTLS the following representation of behaviours has proved convenient, see [11] and [12]. An alternative to the polynomial representation ar is given in terms of state variables x and driving variables v by means of the equations

$$x(t+1) = Ax(t) + Bv(t) \quad (6)$$

$$w(t) = Cx(t) + Dv(t). \quad (7)$$

This can be interpreted as a dynamic factor model, where the observed variables w are generated by the factors x that evolve over time and with unobserved driving forces v . A behaviour of complexity (m, n) can be represented by m driving variables and n state variables, and not by a smaller number of these latent variables. Such representations are called minimal. The minimal representation of a behaviour is highly non-unique. In particular one can choose a so-called minimal isometric representation, so that the $(n+q) \times (n+m)$ matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \quad (8)$$

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

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^T \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I_n & O \\ O & I_m \end{pmatrix} \quad (9)$$

So these are isometric representations, but note that minimality is not required. For $\theta \in \Theta$ we denote by (θ) the behaviour corresponding to θ , that is, all the trajectories w that can be generated by the equations state1, state2 for these values of the parameters. This parametrization is not injective, due to the non-uniqueness of isometric representations. The image of this parametrization consists of behaviours with complexity (m, n') with $n' \leq n$, and $n' = n$ if θ is minimal. The behavioural error is defined by $e(\theta) := e((\theta))$ given in erbehav, and the GTLS error by $e_N(\theta) := e_N((\theta))$ as in GTLS. More explicit expressions in terms of the parameters can be obtained. For instance, $e(\theta)$ is given by erbehav where $P(e^{i\lambda}) = G(e^{i\lambda})G^T(e^{-i\lambda})$ with $G(e^{i\lambda}) = D + C(e^{i\lambda}I - A)^{-1}B$ the transfer function from v to w in the state model state1, state2. The computation of $e_N(\theta)$ by means of a state model is described in [11].

Apart from the parameter set Θ we will also consider the subsets Θ_ρ , with $0 < \rho \leq 1$, defined by

$$\Theta_\rho = \{\theta \in \Theta; \lambda_{max}(A) \leq \rho\} \quad (10)$$

where $\lambda_{max}(A)$ denotes the maximum modulus of the eigenvalues of the matrix A . Because of the isometry condition there holds that $\Theta_1 = \Theta$. The following properties of the parametrization state1, state2, parset are proven in [7], where $\|\cdot\|_\rho$ denotes the supremum norm over Θ_ρ , that is, it is defined by $\|e_N - e\|_\rho := \sup_{\theta \in \Theta_\rho} |e_N(\theta) - e(\theta)|$.

1. Θ and Θ_ρ are compact;
2. e is continuous on Θ ;

3. for fixed N and $\theta_k \rightarrow \theta$ there holds $\limsup_{k \rightarrow \infty} e_N(\theta_k) \leq e_N(\theta)$;
4. e_N is discontinuous only at non-minimal $\theta \in \Theta$;
5. for every $\rho < 1$ holds $\|e_N - e\|_\rho \rightarrow 0$ almost surely if $N \rightarrow \infty$.

5 Consistency

The results in the foregoing section imply the following consistency result in terms of the system parameters. For fixed complexity (m, n) we denote by $\Theta_\rho^* = \operatorname{argmin}_{\theta \in \Theta_\rho} e(\theta)$ the set of optimal parameters and by $\Theta_{\rho, N}^* = \operatorname{argmin}_{\theta \in \Theta_\rho} e_N(\theta)$ the set of parameters of GTLS models for the observed data, and $\|\cdot\|$ denotes the Euclidean norm on Θ as subset of $(n+q)(n+m)$. The following result is proved in [7].

1. Θ_ρ^* is not empty;
2. $\sup_{\theta_N \in \Theta_{\rho, N}^*} \inf_{\theta \in \Theta_\rho^*} \|\theta_N - \theta\| \rightarrow 0$ almost surely in case $N \rightarrow \infty$.

This means that the parameters of optimal models, obtained by the GTLS procedure applied to finite observed data, converge to the parameters of models that are optimal for the data generating process. Similar results can be obtained on the level of behaviours, that is, independent of the chosen parametrization. For simplicity we will only state the result for the case that the optimal behaviour is unique, which holds generically true. Here genericity of data generating processes is defined in terms of the spectrum $\S(e^{i\lambda})$ of the process w . Let S be the set of $q \times q$ spectra that are bounded on the unit circle, and let a metric on S be defined by $d(\S_1, \S_2) = \sup_{\varphi \in [-\pi, \pi]} \lambda_{\max} \{\S_1(e^{i\varphi}) - \S_2(e^{i\varphi})\}$. Then a subset $S' \subset S$ is called generic if it contains a subset $S'' \subset S'$ that is open and dense in S .

For generic data generating process w , the behaviour of complexity (m, n) minimizing the error behaves and is unique, and the same holds true almost surely for the minimization of GTLS.

In order to formulate consistency in terms of behaviours we use the following so-called gap metric

$$d_{(1,2)} = \|G_1 G_1^* - G_2 G_2^*\| \quad (11)$$

where $\|\cdot\|$ denotes the supremum norm on the unit circle, $G_j(e^{i\lambda}) = D_j + C_j(e^{i\lambda}I - A_j)^{-1}B_j$ with $\theta_j = (A_j, B_j, C_j, D_j)$ an isometric representation of j , and with $G_j^*(e^{i\lambda}) = G_j^T(e^{-i\lambda})$, so that $G_j G_j^*$ is the transfer function corresponding to the operator of orthogonal

projection onto the behaviour j , $j = 1, 2$. In the following we assume that the complexity (m, n) is fixed and that a bound $\rho < 1$ has been chosen. By (m, n, ρ) we denote the optimal behaviour over Θ_ρ for the data generating process, and by ${}_N(m, n, \rho)$ the GTLS behaviour over Θ_ρ . We assume that these behaviours are unique, which because of the foregoing proposition is the generic situation. In this case $(m, n, \rho) = (\theta)$ with $\theta \in \Theta_\rho^*$, and ${}_N(m, n, \rho) = (\theta_N)$ with $\theta_N \in \Theta_{\rho, N}^*$. The following result is proved in [7].

Under the above conditions, for fixed complexity (m, n) and $\rho < 1$, there holds $d((m, n, \rho), {}_N(m, n, \rho)) \rightarrow 0$ almost surely for $N \rightarrow \infty$.

So the optimal behaviour identified from observed data converges to the optimal behaviour for the data generating process if the number of observations tends to infinity. This provides an answer to the consistency question formulated in cons.

6 Conclusion

In this paper the GTLS method for linear system identification has been investigated within a stochastic framework. The object of interest is the system behaviour. This approach treats all variables in a symmetric way and the basic system properties are expressed independent of parametrization. The central result concerns the consistency of GTLS, in the sense that the identified model converges to an optimal approximation of the data generating process. Here the complexity of the models is fixed. The main restriction is that the maximum modulus of the eigenvalues of the state transition matrix is bounded a priori by a fixed number $\rho < 1$, while the full parameter set has $\rho \leq 1$ (in fact, representations with $\rho = 1$ are not minimal and hence correspond to systems of lower complexity). This restriction is very helpful for proving the consistency result, details of which are given in [7]. We conjecture that consistency without this restriction, that is, on the full class of behaviours of fixed complexity, does also hold true. This is a topic of further research.

References

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

- [2] M. Deistler and W. Scherrer, Identification of linear systems from noisy data. In D. Brillinger, ed., *New Directions in Time Series Analysis*, part II, IMA vol. 46, Springer, 1992, pp. 21-42.
- [3] E.J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. Wiley, 1988.
- [4] C. Heij, *Deterministic Identification of Dynamical Systems*. Springer, 1989.
- [5] C. Heij and B. Roorda, Noise minimization in factor models. In M. Blanke and T. Söderström, eds., *Preprints of the 10th IFAC Symposium on System Identification*, Vol.3, 1994, pp. 203-206.
- [6] C. Heij and W. Scherrer, On the consistency of identification by dynamic factor models. *Proceedings 33rd CDC*, Florida, 1994, pp. 2880-2885.
- [7] C. Heij and W. Scherrer, Consistency of GTLS modelling. *Mimeo*, 1995.
- [8] C. Heij, W. Scherrer and M. Deistler, System identification by dynamic factor models. *Report 9501/A*, Econometric Institute, Erasmus University Rotterdam, The Netherlands, 1995. Submitted for publication.
- [9] L. Ljung, *System Identification : Theory for the User*. Prentice-Hall, 1987.
- [10] B. Roorda, *Global Total Least Squares*. Ph.D. Thesis, Tinbergen Institute Research Series, vol. 88, 1995.
- [11] B. Roorda, Algorithms for global total least squares modelling of finite multivariable time series. *Automatica* 31, 1995, pp. 391-404.
- [12] B. Roorda and C. Heij, Global total least squares modelling of multivariable time series. *IEEE Transactions on Automatic Control* 40, 1995, pp. 50-63.
- [13] J.C. Willems, From time series to linear system (three parts). *Automatica* 22,23, 1986, 1987, pp. 561-580, 675-694, 87-115.
- [14] J.C. Willems, Paradigms and puzzles in the theory of dynamical systems. *IEEE Transactions on Automatic Control* 36, 1991, pp. 259-294.