Estimation in a linear multivariate measurement error model with a change point in the data

A. Kukush[1,*], I. Markovsky[2], and S. Van Huffel[3]

1 — Kiev National Taras Shevchenko University, Vladimirskaya st. 64, 01033, Kiev, Ukraine

2 — School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK

3 — ESAT, SCD - SISTA, K. U. Leuven, Kasteelpark Arenberg 10, B - 3001 Leuven, Belgium

Abstract

A linear multivariate measurement error model $AX = B$ is considered. The errors in $\begin{bmatrix} A & B \end{bmatrix}$ are row-wise finite dependent, and within each row, the errors may be correlated. Some of the columns may be observed without errors, and in addition the error covariance matrix may differ from row to row. The columns of the error matrix are united into two uncorrelated blocks, and in each block, the total covariance structure is supposed to be known up to a corresponding scalar factor. Moreover the row data are clustered into two groups, according to the behavior of the rows of true A matrix. The change point is unknown and estimated in the paper. After that, based on the method of corrected objective function, strongly consistent estimators of the scalar factors and X are constructed, as the numbers of rows in the clusters tend to infinity. Since Toeplitz/Hankel structure is allowed, the results are applicable to system identification, with a change point in the input data.

Key words: Linear errors-in-variables model; Corrected objective function; Clustering; Dynamic errors-in-variables model; Consistent estimator.

1 Introduction

We deal with a multivariate multiple errors-in-variables (EIV) model. Our assumptions are rather general and comprise both the element-wise weighted total least squares problem, see Kukush and van Huffel (2004), and the structured total least squares problem, see Kukush et al. (2005b). A key condition in these papers is that the noise covariance structure is known up to a scaling factor. One can argue that such a knowledge is again respective in practice.

The EIV models with two or more unknown noise parameters, however, are non-identifiable by second order methods. This problem of non-identifiability is well known in the context of the Frisch scheme, see Frisch (1934) and De Moor (1988). A similar negative result for dynamical systems is first proven in Anderson (1985).

Various additional assumptions can be imposed in order to make the EIV estimation problem with unknown noise covariance structure identifiable. An overview of methods for EIV system identification is given in Söderström et al. (2002).

In this paper, we assume that the errors matrix is partitioned into two uncorrelated blocks, and in each block, the total covariance structure is known up to a corresponding scalar factor. The condition about the two unknown scalar factors is common, e.g., such situation arises in dynamic case where the input and output matrices A and B are stochastically independent and their covariance structures are known up to two scalars, λ_A and λ_B, say. Similar problems arise in static cases, see Cirrincione et al. (2001). Zheng (1998) proposed bias-estimated least-squares estimated algorithms for such dynamic problems. See overview of different approaches in Agüero and Goodwin (2006). Zheng (1998) assumed the true input process to be stationary with rational spectral density, while the input and output errors to be white noises. In the present paper we allow the true input and measurement errors to have similar covariance structure, which causes non-identifiability of the system.

*Corresponding author. Tel.: +380 44 2590591; Fax: +380 44 2590392
Email address: alexander_kukush@univ.kiev.ua (Alexander Kukush)

We show that the new assumption enables to derive the consistent parameter and noise variance estimates. Namely, we assume that the true row data are clustered into two groups. This corresponds to a change point in case of a dynamical EIV model. The first attempt to use clustering in a linear measurement error model with unknown noise variances was made by Wald (1940), where the scalar case is treated. The idea used in the paper was to cluster the data into two groups and draw a straight line through the means of the two groups. The clustering criterion is that the means are asymptotically separated from each other. Only the first empirical moment is used in the construction of the estimator for the slope and the intercept.

We further develop and extend the approach of Wald (1940). Our clustering assumption is based on the second moment of the rows of true matrix. In the scalar case, it is possible that the means in the groups are close to each other but our clustering condition still holds. We allow groups with the same mean but with different dispersions. In a scalar model considered by Wald (1940), our resulting estimator is different from Wald's estimator since we utilize the second empirical moment also.

The proposed estimation procedure consists of three steps: 1) cluster the data into two groups, 2) estimate the noise variances λ_A and λ_B, and 3) estimate the parameter of interest using the noise variance estimates. The optimization procedure at the first step is rather simple and based on the second empirical moment. The optimization problem in the second step is, in general, nonconvex and nonsmooth. In our simulation examples, we apply general optimization methods for its solution, e.g., the simplex method, see Nelder and Mead (1965). The third step involves an eigenvalue decomposition of a symmetric matrix, so it is computationally inexpensive.

The assumptions that the data can be clustered means that the true input changes its character, while the noise properties remain the same. This assumption can be viewed equivalently as having a set of data records from experiments with different true inputs. Such an assumption is certainly restrictive. The proposed method is not applicable to the problems where the inputs are stationary, which is a typical assumption in most of the earlier works on EIV system identification. The situation is similar to Wald's estimator in the present case. Madansky (1959) noted that when clusters are given a priori, Wald's method is an instrumental variables method for estimating the slope, but this is not the case when the clusters are chosen from the data. Indeed, Pakes (1982) shows that Wald's estimator is inconsistent when there is no change point in the data and clusters are chosen by the data in the way recommended by Wald.

We mention here two papers which are closely related to the present work. In the technical report Kukush et al. (2005a), a similar approach is used for two inputs two outputs systems, which means that the change point in the input data is known. In Markovsky et al. (2006) another estimator is proposed in the presence of two clusters. That estimator is easier to compute but its asymptotic properties are unclear.

We use the following notations. $\|A\|$ is the Frobenius norm of the matrix A. I_p denotes a unit matrix of size p. For a symmetric matrix C, $\mu_1(C) \leq \ldots \leq \mu_p(C)$ are the p smallest eigenvalues of C.

The paper is organized as follows. Section 2 presents a general $AX = B$ model, without clustering condition and with rather mild assumptions on the error terms. Here we use the method of corrected objective function to derive an estimator of X in case of known scalar factors λ_1^0 and λ_2^0 and make a preliminary attempt to derive an objective function for λ_1^0, λ_2^0 when they are unknown. In Section 3, we introduce a model with two clusters, and in Section 4, we estimate the change point consistently. Next in Section 5, we utilize the clustering idea to introduce the final objective function for the scalar factors and state the consistency result. In Section 6 the consistent estimator for X is proposed. A simulation experiment for the proposed study is discussed in Section 7 and Section 8 presents some conclusion. The proofs are given in Appendices.

2 General model without clustering

2.1 General assumptions

Consider the model

$$AX = B, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times p}$, and $B \in \mathbb{R}^{m\times p}$. Equivalently, the model is written as

$$DX_{\text{ext}} = 0,$$

where

$$D := \begin{bmatrix} A & B \end{bmatrix} \quad \text{and} \quad X_{\text{ext}} := \begin{bmatrix} X \\ -I_p \end{bmatrix}. \tag{2}$$

The model (1) means the following. For true values, we have

$$\bar{A}X = \bar{B}, \tag{3}$$

where X is nonrandom matrix. We observe

$$A = \bar{A} + \tilde{A} \quad \text{and} \quad B = \bar{B} + \tilde{B}. \tag{4}$$

Here $\tilde{A}$ and $\tilde{B}$ are random matrices, which are stochastically independent of $\bar{A}$. Alternatively, we can write

$$D = \bar{D} + \tilde{D}, \quad \bar{D}X_{\text{ext}} = 0.$$

Here $\bar{D} := \begin{bmatrix} \bar{A} & \bar{B} \end{bmatrix}$ and $\tilde{D} := \begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix}$. We want to estimate X with fixed n and p and increasing m.

Let $\tilde{d}_{ij}$, $1 \le i \le m$, $1 \le j \le n+p$, be the entries of $\tilde{D}$, and $\tilde{D}^\top = \begin{bmatrix} \tilde{d}_1 & \dots & \tilde{d}_m \end{bmatrix}$, similar notation will be used for the rows of other matrices. Concerning the errors $\tilde{d}_{ij}$, we assume the following.

(i). $\mathbf{E}\,\tilde{d}_i = 0$, for all i.
Hereafter $\mathbf{E}$ denotes the expectation of a random variable, vector, or matrix.

(ii). There exists $\delta > 0$, such that

$$\sup_{i \ge 1} \mathbf{E}\,\|\tilde{d}_i\|^{4+\delta} < \infty.$$

(iii). The sequence of random vectors $\{\tilde{d}_i,\ i \ge 1\}$ is finite dependent.

This means that there exists a $q \ge 0$ such that for each $k \ge 1$, the two sequences

$$\{\tilde{d}_1, \dots, \tilde{d}_k\} \quad \text{and} \quad \{\tilde{d}_{k+q+1}, \tilde{d}_{k+q+2}, \dots\}$$

are independent from each other.

Note 1. It is possible to weaken the condition (iii) by assuming that $\{\tilde{d}_i,\ i \ge 1\}$ are weakly dependent with appropriate condition on the mixing coefficients. Then one can use Rosenthal moment inequality for weakly dependent random variables, see Doukhan (1994).

(iv). There exists n_1, $1 \le n_1 \le n+p-1$, such that $\mathbf{E}\,\tilde{d}_{ij}\tilde{d}_{ik} = 0$, for each $i \ge 1$, $1 \le j \le n_1$, $n_1+1 \le k \le n+p$.

This means that the error matrix $\tilde{D}$ can be partitioned as $\tilde{D} = \begin{bmatrix} \tilde{D}_1 & \tilde{D}_2 \end{bmatrix}$, $\tilde{D}_1 \in \mathbb{R}^{m \times n_1}$, into two blocks with

$$\mathbf{E}\,\tilde{D}_1^\top \tilde{D}_2 = 0.$$

(v). $\mathbf{E}\,\tilde{D}_k^\top \tilde{D}_k = \lambda_k^0 W_k$, $k = 1, 2$, where W_k are known positive semidefinite matrices and λ_k^0 are unknown positive scalars.

One may recall that a symmetric C is said to be positive semidefinite if its quadratic form is nonnegative.

In this paper the true matrix $\bar{A} = \begin{bmatrix} \bar{a}_1 & \cdots & \bar{a}_m \end{bmatrix}^\top$ is assumed to be random.

(vi). Random vectors $\{\bar{a}_i, i = 1, 2, \dots\}$ are identically distributed and form a finite dependent sequence, with finite second moments.

Summarizing we can say that D is observed with known W_1 and W_2. Our goal is to estimate X consistently, as $m \to \infty$.

2.2 Derivation of the score function

Suppose first that λ_1^0 and λ_2^0 are known. The question now is how to estimate X by the corrected objective function method, see Kukush and Zwanzing (2001) for the concerned method. It is closely related to the method of Corrected Score, see Carroll et al. (1995), Chapter 4.

The least squares objective function is

$$Q_{\mathrm{ls}}(\bar{D};Z) := \|\bar{D}Z\|^2, \quad Z \in \mathbb{R}^{(n+p)\times p},$$

which can also be represented as

$$Q_{\mathrm{ls}}(\bar{D};Z) = \mathbf{tr}(Z^\top \Psi_{\mathrm{ls}}(\bar{D})Z),$$

where

$$\Psi_{\mathrm{ls}}(\bar{D}) := \bar{D}^\top \bar{D}.$$

By the method of corrected objective function, we need to construct $Q_c(D;Z)$, such that

$$\mathbf{E}[Q_c(D;Z) \mid \bar{D}] = Q_{\mathrm{ls}}(\bar{D};Z) \text{ , for all } Z,$$

which is possible under known λ_k^0 and W_k, $k = 1,2$, defined as in conditions (iv) – (v). Let

$$\Psi_c(D) = D^\top D - \begin{bmatrix} \lambda_1^0 W_1 & 0 \\ 0 & \lambda_2^0 W_2 \end{bmatrix}.$$

Then

$$Q_c(D;Z) = \mathbf{tr}(Z^\top \Psi_c(D) Z).$$

We minimize this objective function, subject to

$$Z^\top Z = I_p.$$

This is a convex optimization problem on a Grassman manifold, see, e.g., Edelman et al. (1998). For the solution $\hat{Z} =: \begin{bmatrix} \hat{Z}_1 \\ \hat{Z}_2 \end{bmatrix}$, $\hat{Z}_2 \in \mathbb{R}^{p\times p}$, the estimator of X would be

$$\hat{X} := -\hat{Z}_1 (\hat{Z}_2)^{-1},$$

provided $\hat{Z}_2$ is nonsingular. Let $\mu_i := \mu_i\big(\Psi_c(D)\big)$, $i = 1,2,\ldots,n+p$, be the ordered eigenvalues of $\Psi_c(D)$ and $\{\varphi_i,\ i = 1,2,\ldots,n+p\}$ be the corresponding orthonormal eigenbasis. If $\mu_{p+1} > \mu_p$, then $\hat{Z} = [\varphi_1\ \ldots\ \varphi_p]$, and $\hat{X}$ does not depend on the choice of the eigenbasis.

Lemma 2.1. *Under the assumptions (i) to (iv), (vi)and assuming (v) as well as λ_k^0 to be known, then*

$$\left\| \frac{1}{m}\Psi_c(D) - \frac{1}{m}\Psi_{\mathrm{ls}}(\bar{D}) \right\| \to 0, \quad as \quad m \to \infty, \quad a.s.$$

Proof. The proof is an easy application of a matrix version of the Rosental moment inequality, see Kukush *et al.* (2005b), Lemma 2(b), and can be obtained as per the lines of the proof of Lemma 3.1 of technical report Kukush *et al.* (2005a). □

2.3 Constructing the cost function under unknown λ_k^0

For $\lambda_1, \lambda_2 \geq 0$, define

$$\Psi_c(D;\lambda_1,\lambda_2) = \Psi_c(\lambda_1,\lambda_2) := D^\top D - \begin{bmatrix} \lambda_1 W_1 & 0 \\ 0 & \lambda_2 W_2 \end{bmatrix},$$

and

$$\Psi_{\mathrm{ls}}(\bar{D};\lambda_1,\lambda_2) = \Psi_{\mathrm{ls}}(\lambda_1,\lambda_2) := \bar{D}^\top \bar{D} - \begin{bmatrix} (\lambda_1 - \lambda_1^0)W_1 & 0 \\ 0 & (\lambda_2 - \lambda_2^0)W_2 \end{bmatrix}.$$

By Lemma 2.1,

$$\left\| \frac{1}{m}\Psi_c(\lambda_1,\lambda_2) - \frac{1}{m}\Psi_{\mathrm{ls}}(\lambda_1,\lambda_2) \right\| \to 0, \quad \text{as} \quad m \to \infty, \quad \text{a.s.}$$

We study the properties of the approximating matrix $\frac{1}{m}\Psi_{\mathrm{ls}}(\lambda_1,\lambda_2)$.

(vii). $\mathbf{E}\,\bar{a}_1\bar{a}_1^\top$ is positive definite.

Lemma 2.2. *Under (3) and condition (vi) – (vii), we have a.s., as $m \to \infty$:*

$$\frac{1}{m}\bar{D}^\top \bar{D} \to \Psi_\infty := \begin{bmatrix} I_n \\ X^\top \end{bmatrix} \mathbf{E}\, \bar{a}_1 \bar{a}_1^\top \begin{bmatrix} I_n & X \end{bmatrix},$$

and $\mu_{p+1}(\Psi_\infty) > 0$.

(Here μ_{p+1} is the $(p+1)$th smallest eigenvalue.)

Proof. The proof is straightforward and it is not given here. □

Thus for large m, we derive the approximate equality

$$\frac{1}{m}\Psi_c(\lambda_1, \lambda_2) \approx \frac{1}{m}\Psi_{\rm ls}(\lambda_1, \lambda_2),$$

and for this approximate matrix, we have

$$\mu_i\big(\Psi_{\rm ls}(\lambda_1^0, \lambda_2^0)\big) = 0, \quad \text{for all} \quad 1 \le i \le p,$$

and $\mu_{p+1}\big(\Psi_{\rm ls}(\lambda_1^0, \lambda_2^0)\big)$ is positive and separated from 0. Moreover

$$L_p(\lambda_1^0, \lambda_2^0) = \operatorname{span}(z_1, \dots, z_p),$$

where $L_p(\lambda_1^0, \lambda_2^0)$ is the kernel of $\Psi_{\rm ls}(\lambda_1^0, \lambda_2^0)$ and $\begin{bmatrix} z_1 & \dots & z_p \end{bmatrix} := \begin{bmatrix} X \\ -I_p \end{bmatrix}$.

In order to estimate λ_1^0, λ_2^0, we could use the cost function

$$Q(\lambda_1, \lambda_2) := \sum_{i=1}^{p} \mu_i^2 \left(\frac{1}{m}\Psi_c(\lambda_1, \lambda_2) \right)$$

and minimize it for $\lambda_1, \lambda_2 \ge 0$. Unfortunately this does not yield a consistent estimator of $(\lambda_1^0, \lambda_2^0)$ since for the approximating matrix-valued function $\Psi_{\rm ls}(\lambda_1, \lambda_2)/m$ there could be other values λ_1^*, λ_2^*, separated from λ_1^0 and λ_2^0, with

$$\mu_i\big(\Psi_{\rm ls}(\lambda_1^*, \lambda_2^*)\big) = 0, \quad \text{for all} \quad 1 \le i \le p.$$

Therefore the minimum point of $Q(\lambda_1, \lambda_2)$ need not converge in probability to $(\lambda_1^0, \lambda_2^0)$, as $m \to \infty$.

Note 2. The function $Q(\lambda_1, \lambda_2)$ is continuous at the domain $\lambda_1, \lambda_2 \ge 0$.

3 Model with two clusters

Once again consider the model (1). We observe

$$A = \bar{A} + \tilde{A}, \quad B = \bar{B} + \tilde{B}$$

with

$$\bar{A}X = \bar{B}.$$

Here $X \in \mathbb{R}^{n \times p}$ is nonrandom matrix to be estimated, $A \in \mathbb{R}^{m \times n}$, and $B \in \mathbb{R}^{m \times p}$. The number of rows $m = m(t)$, where $t = 1, 2, 3, \dots$ stands for the number of experiment and $m(t) \to \infty$, as $t \to \infty$. The dimensions n and p are fixed.

Let $\{u_i,\ i \ge 1\}$ and $\{v_i,\ i \ge 1\}$ be two mutually independent sequences of $\mathbb{R}^{n \times 1}$ random vectors;

$u_i \stackrel{d}{=} u$, $i \ge 1$; $v_i \stackrel{d}{=} v$, $i \ge 1$, which means that $\{u_i\}$ have same distribution, and $\{v_i\}$ have (another) common distribution. We suppose that both sequences $\{u_i\}$ and $\{v_i\}$ are finite dependent.

Now, we need that for each $t \ge 1$

$$\bar{A}^\top = \bar{A}^\top(t) = \begin{bmatrix} u_1 & \dots & u_{m_1(t)} & f_1(t) & \dots & f_q(t) & v_1 & \dots & v_{m_2(t)} \end{bmatrix}.$$

Here $m_1 = m_1(t)$ is a change point, $q \ge 0$ is fixed and does not depend of t, and

$$m(t) = m_1(t) + q + m_2(t).$$

Moreover, we suppose that $m_1(t) \geq q$, $m_2(t) \geq q$, and random vectors $f_i(t)$ satisfy

$$\| f_i(t)\| \leq const \cdot \left(\sum_{i=m_1(t)-q+1}^{m_1(t)} \|u_i\| + \sum_{i=1}^{q} \|v_i\| \right), \ i = 1, \ldots, q. \tag{5}$$

Thus actually we have certain transition regime for the rows of $\bar{A}(t)$ with numbers from $m_1(t)+1$ till $m_1(t)+q$, after that we have totally different distribution of rows. We allow $q = 0$ which means that the change in behaviour of the rows in $\bar{A}$ happens immediately after the change point $m_1(t)$.

Concerning the clusters assume the following.

(cl_1) $m_1(t)/m(t) \to r$, as $t \to \infty, 0 < r_1 \leq r \leq r_2 < 1$,

and the bounds r_1 and r_2 are known.

(cl_2) For certain $\delta > 0$, $\mathbf{E}\,\|u\|^{2+\delta} < \infty$, $\mathbf{E}\,\|v\|^{2+\delta} < \infty$, and

$$\sigma_1(\mathbf{E}\,uu^\top - \mathbf{E}\,vv^\top) > 0, \tag{6}$$

where $\sigma_1(C)$ is the smallest singular value of the symmetric matrix C.

Inequality (6) is crucial clustering assumption. It holds, e.g., if either $\mathbf{Var}(u) = \mathbf{Var}(v)$ and $\mathbf{E}\,u$, $\mathbf{E}\,v$ are linearly independent (as considered in Wald (1940)), or $\mathbf{E}\,u = \mathbf{E}\,v$ and $\mathbf{Var}(u) - \mathbf{Var}(v)$ is nonsingular.

Let $D = \begin{bmatrix} A & B \end{bmatrix}$, $\bar{D} = \begin{bmatrix} \bar{A} & \bar{B} \end{bmatrix}$, and $\tilde{D} = \begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix}$, where all the matrices depend of t. We assume that similarly to the structure of the matrix $\bar{A}(t)$,

$$\tilde{D}^\top = \begin{bmatrix} \tilde{d}_1 & \ldots & \tilde{d}_{m_1(t)} & \tilde{e}_1 & \ldots & \tilde{e}_q & \tilde{h}_1 & \ldots & \tilde{h}_{m_2(t)} \end{bmatrix}.$$

Here $\tilde{d}_i$, $i \geq 1$, and $\tilde{h}_i$, $i \geq 1$, are two mutually independent sequences of $\mathbb{R}^{(n+p)\times 1}$ random vectors, and random vectors $\tilde{e}_1(t), \ldots, \tilde{e}_q(t)$ satisfy inequality

$$\|\tilde{e}_i(t)\| \leq const \cdot \left(\sum_{i=m_1(t)-q+1}^{m_1(t)} \|\tilde{d}_i\| + \sum_{i=1}^{q} \|\tilde{h}_i\| \right), \ i = 1, \ldots, q. \tag{7}$$

We assume that $\tilde{D}(t)$ is independent of $\bar{A}(t)$ for each $t \geq 1$. Moreover concerning the errors we demand the following.

(a). $\mathbf{E}\,\tilde{d}_i = 0$, $\mathbf{E}\,\tilde{h}_i = 0$, $\mathbf{E}\,\tilde{e}_k(t) = 0$ for all $i \geq 0$ and $k = 1, \ldots, q$, $t \geq 1$.

(b). There exists $\delta > 0$, such that

$$\sup_{i\geq 1} \mathbf{E}\,\|\tilde{d}_i\|^{4+\delta} < \infty, \quad \sup_{i\geq 1} \mathbf{E}\,\|\tilde{h}_i\|^{4+\delta} < \infty.$$

(c). The sequences of random vectors $\{\tilde{d}_i,\ i \geq 1\}$ and $\{\tilde{h}_i,\ i \geq 1\}$ are finite dependent.

(d). The errors matrix $\tilde{D} = \tilde{D}(t)$ can be partitioned as

$$\tilde{D} = \begin{bmatrix} \tilde{D}_1 & \tilde{D}_2 \end{bmatrix}, \ \tilde{D}_1 \in \mathbb{R}^{m(t)\times n_1}$$

into two blocks satisfying the condition:

$$\mathbf{E}\,\tilde{D}_1^\top \tilde{D}_2 = 0.$$

(e). $\mathbf{E}\,\tilde{D}_i^\top \tilde{D}_i = \lambda_i^0 W_i$, $i = 1, 2$, where W_i are known positive semidefinite matrices depending of t, and λ_i^0, are unknown positive scalars which do not depend of t. Introduce a partition of the matrix (2)

$$X_{\text{ext}} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad X_1 \in \mathbb{R}^{n_1\times p}, \quad X_2 \in \mathbb{R}^{n_2\times p}.$$

(f). $\liminf\limits_{t\to\infty}\ \mathbf{tr}\left(X_j^\top (W_j/m_j(t)) X_j\right) > 0$, for $j = 1, 2$.

(g).

$$\frac{1}{m_1(t)}\sum_{i=1}^{m_1(t)}\mathbf{E}\,\tilde{d}_i\bar{d}_i^{\top}-\frac{1}{m_2(t)}\sum_{i=1}^{m_2(t)}\mathbf{E}\,\tilde{h}_i\bar{h}_i^{\top}\to 0,\ \text{as } t\to\infty.$$

We demand also that the signal component of the data does not degenerate:

(h). $\mathbf{E}\,uu^{\top}+\mathbf{E}\,vv^{\top}$ is positive definite.

4 Estimation of the change point

Define an objective function

$$F(m_1)=\sigma_1\left(\frac{1}{m_1}\sum_{i=1}^{m_1}a_ia_i^{\top}-\frac{1}{m-m_1}\sum_{i=m_1+1}^{m}a_ia_i^{\top}\right),\ r_1m\le m_1\le r_2m. \tag{8}$$

Here $A^{\top}=:\begin{bmatrix}a_1 & a_2 & \dots & a_m\end{bmatrix}$. Remember that r_1 and r_2 enter the condition (cl_1).

Define the estimator $\hat{m_1}$ for $m_1(t)$ as a Borel measurable discrete function of the observation matrix A that satisfies

$$F(\hat{m}_1)=\max_{r_1m\le m_1\le r_2m}F(m_1). \tag{9}$$

The next statement is a result for the consistency of ratio $\hat{m}_1/m(t)$, as the number of experiment t is increasing. Remember that the limit ratio r is introduced in condition (cl_1).

Theorem 4.1. *Under the conditions (cl_1), (cl_2), and (a) to (c), $\hat{m}_1/m(t)\to r$, as $t\to\infty$, a.s.*

The proofs of all the statements are given in Appendix.

5 Estimation of two scale factors

After the estimation of the change points, the rows of D matrix can be partitioned into two parts,

$$D=\begin{bmatrix}D(1)\\D(2)\end{bmatrix},\ D(1)\in\mathbb{R}^{\hat{m}_1\times(n+p)},$$

and respectively

$$\tilde{D}=\begin{bmatrix}\tilde{D}(1)\\\tilde{D}(2)\end{bmatrix},\ \tilde{D}(1)\in\mathbb{R}^{\hat{m}_1\times(n+p)},$$

and similarly for the true values $\bar{D}$. From the condition (d), we have further partition

$$\tilde{D}=\begin{bmatrix}\tilde{D}_1(1) & \tilde{D}_2(1)\\\tilde{D}_1(2) & \tilde{D}_2(2)\end{bmatrix},\tilde{D}_1(1)\in\mathbb{R}^{\hat{m}_1\times n_1}.$$

Thus

$$\tilde{D}_1=\begin{bmatrix}\tilde{D}_1(1)\\\tilde{D}_1(2)\end{bmatrix},\tilde{D}_2=\begin{bmatrix}\tilde{D}_2(1)\\\tilde{D}_2(2)\end{bmatrix}.$$

From condition (e) we have for certain matrices $W_1(k)$:

$$\mathbf{E}[\tilde{D}_1^{\top}(k)\tilde{D}_1(k)|\hat{m}_1]=\lambda_1^0W_1(k),k=1,2.$$

Thus the matrix W_1 in (e) satisfies

$$W_1=\begin{bmatrix}W_1(1) & V_1\\V_1^{\top} & W_1(2)\end{bmatrix}.$$

Similarly

$$\mathbf{E}[\tilde{D}_2^{\top}(k)\tilde{D}_2(k)|\hat{m}_1]=\lambda_2^0W_2(k),k=1,2,$$

and

$$W_2 = \begin{bmatrix} W_2(1) & V_2 \\ V_2^\top & W_2(2) \end{bmatrix}.$$

Let for $\lambda := (\lambda_1, \lambda_2)^\top \in [0,\infty) \times [0,\infty)$,

$$\Psi_c^{(k)}(\lambda) = D^\top(k)D(k) - \begin{bmatrix} \lambda_1 W_1(k) & 0 \\ 0 & \lambda_2 W_2(k) \end{bmatrix},$$

and $\mu_{1k}(\lambda) \le \mu_{2k}(\lambda) \le \ldots \le \mu_{pk}(\lambda)$ be the p smallest eigenvalues of $\Psi_c^{(k)}(\lambda)$ with the corresponding orthonormal eigenvectors $f_{1k}(\lambda), \ldots, f_{pk}(\lambda)$. In case of multiple eigenvalues the $f_{ik}(\lambda)$ are not uniquely defined. Then we define them in such a way that they are Borel measurable function of $D(k)$ and λ.

Let $L_{pk}(\lambda)$ be the span of $f_{1k}(\lambda), \ldots, f_{pk}(\lambda)$. Define an objective function

$$Q_c(\lambda) = \sum_{k=1}^{2} \sum_{i=1}^{p} \mu_{ik}^2(\lambda) + c\|\sin\Theta(\lambda)\|^2. \tag{10}$$

Here $c > 0$ is a fixed constant and $\Theta(\lambda)$ is a diagonal matrix of canonical angles between $L_{p1}(\lambda), L_{p2}(\lambda)$, and $sin\,\Theta(\lambda)$ is defined as the diagonal matrix with diagonal elements the sines of these angles, see Stewart and Sun (1990), p.43, Corollary 5.4. Actually the sines of the canonical angles between two subspaces U_1 and U_2 of a real Euclidean space of the same dimension are the nonzero singular values of the matrix $V^\top W$, where the columns of V form an orthonormal basis of the orthogonal complement of U_1 and the columns of W form an orthonormal basis of U_2. We mention that $sin\,\Theta(\lambda)$ is a Borel measurable function of λ and can be discontinuous.

We fix a positive sequence $\{\varepsilon_t, t \ge 1\}$, such that $\varepsilon_t \to 0$, as $t \to \infty$, and define the estimator $\hat{\lambda} = \hat{\lambda}(t)$ as a Borel measurable function of the observations that satisfies the inequality

$$Q_c(\hat{\lambda}) \le \inf_{\lambda_1,\lambda_2 \ge 0} Q_c(\lambda) + \varepsilon_t. \tag{11}$$

Theorem 5.1. *Under the conditions $(cl_1), (cl_2)$, and (a) to (g), $\hat{\lambda} \to \lambda^0 := (\lambda_1^0, \lambda_2^0)^\top$, as $t \to \infty$, a.s.*

6 Final estimator of X

Introduce the matrix

$$\hat{H} := D^\top D - \begin{bmatrix} \hat{\lambda}_1 W_1 & 0 \\ 0 & \hat{\lambda}_2 W_2 \end{bmatrix}.$$

Let $L_p(\hat{H})$ be the subspace spanned by the first p eigenvectors of $\hat{H}$ corresponding to the smallest eigenvalues. Define an estimator $\hat{X}$ by the equality

$$\begin{bmatrix} \hat{X} \\ -I_p \end{bmatrix} = \begin{bmatrix} \hat{z}_1 & \ldots & \hat{z}_p \end{bmatrix}. \tag{12}$$

where

$$L_p(\hat{H}) = \text{span}\,(\hat{z}_1, \ldots, \hat{z}_p). \tag{13}$$

More precisely $\hat{X}$ is a Borel measurable function of D and $\hat{\lambda}$, which satisfies the latter two equalities. It could be not unique since $L_p(\hat{H})$ could be not uniquely defined due to multiple eigenvalues. But we will show that $L_p(\hat{H})$ is unique "eventually", in the following sence.

Definition 6.1. We say that a random event $F = F(t)$ holds eventually if there exists Ω_0, there is $t_0(\omega)$ with the property: for all $t > t_0(\omega)$, the event $F(t)$ holds.

This means that a random event $F(t)$ occurs a.s., starting from a random number t_0.

Theorem 6.1. *Suppose that all the conditions of Theorem 5.1 hold. Assume additionally (h). Then eventually (12)–(13) has a unique solution and $\hat{X} \to X$, as $t \to \infty$, a.s.*

In summary, the proposed estimation procedure has three stages.

(a). Cluster the data by solving the optimization problem (8)–(9).

(b). Compute the noise variance estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ by solving the optimization problem (10)–(11).

(c). Define the estimate $\hat{X}$ by (12)–(13).

7 Simulation example

The simulation example, shown in this section, aims to illustrate the consistency of the proposed estimators for the unknown parameters λ_1^0, λ_2^0, X, and to compare the proposed method with the weighted total least squares method, which assumes that the true noise variane ratio μ^0 is known. Consider the model (3), (4). The error terms in $\tilde{A}$ and $\tilde{B}$ are uncorrelated. The covariance structure of $\tilde{A}$ is known up to a scalar factor λ_1^0 and the covariance structure of $\tilde{B}$ is known up to a scalar factor λ_2^0.

Let $U_{m'}(l,u)$ be a matrix with m' columns, composed of independent and uniformly distributed random variables in the interval $[l,u]$. The true values $\bar{A}$ and X are selected as follows:

$$\bar{A} = \begin{bmatrix} U_{m'}(0,1) \\ U_{m'}(2,4) \end{bmatrix}, \qquad X = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where the two blocks $U_{m'}(0,1)$ and $U_{m'}(2,4)$ are independent and m' is varied from 50 to 750. Correspondingly $\bar{B} := \bar{A}X$. Note that we artificially create two clusters (the first m' and the last m' rows of $\bar{D} = \begin{bmatrix} \bar{A} & \bar{B} \end{bmatrix}$). The measurement errors have zero mean and are independently normally distributed with variances $\lambda_1^0 = 10$ and $\lambda_2^0 = 15$. While minimizing the objective function (10) over λ, we choose the regularization constant $c = 1$ and apply the simplex method, see, e.g., Nelder and Mead (1965), which is a standard method for minimizing a discontinuous objective function.

With this simulation setup, we apply the proposed estimation method and average the results for 100 noise realizations.The average values of the noise variance estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are shown in Figure 1 with solid lines. On the same plots with dashed-dotted lines are shown the average values of the estimates obtained from the weighted total least squares estimator, that takes the true noise variance ratio μ^0 as an input. The dashed lines are the true values λ_1^0 and λ_2^0. Convergence of the average estimates to the true values of the parameters, as the sample size grows, indicates consistency of the estimators. As expected, the weighted total least squares estimates are better than those obtained with the proposed method. The reason is the extra knowledge—true noise variance ratio—that the weighted total least estimator uses. As the sample size grows, however, the difference between the proposed and the weighted total least squares estimates becomes smaller.

Surprisingly, the difference between the estimation accuracy of the parameter of interest X for the proposed and weighted total least squares estimators is much smaller than the one for the parameters λ_1^0 and λ_2^0. Figure 2 shows the average relative estimation error

$$e := \frac{1}{100} \sum_{i=1}^{100} \frac{\|X - \hat{X}^{(i)}\|}{\|X\|},$$

where $\hat{X}^{(i)}$ is the estimate of the parameter X on the i-th repetition of the experiment. For sample sizes m above 150, the two estimators achieve almost the same accuracy (although the accuracy in the λ_1^0 and λ_2^0 estimates is higher for the weighted total least squares estimator).

Finally, we show in Figure 3 the function $f(m')$ that is used for the estimation of the turnover point in the case when the sample size is $m = 1500$. The sharp maximum of $f(m')$ at $m' = 750$ shows that in the example the proposed method allows to detect correctly the turnover point. The example, however, has well separated classes, with difference in both mean and variance. Simulations suggest that equal means of the clusters makes the turnover point estimation harder even if the variances of the clusters still differ. In such cases, the function $f(m')$ has many local maxima.

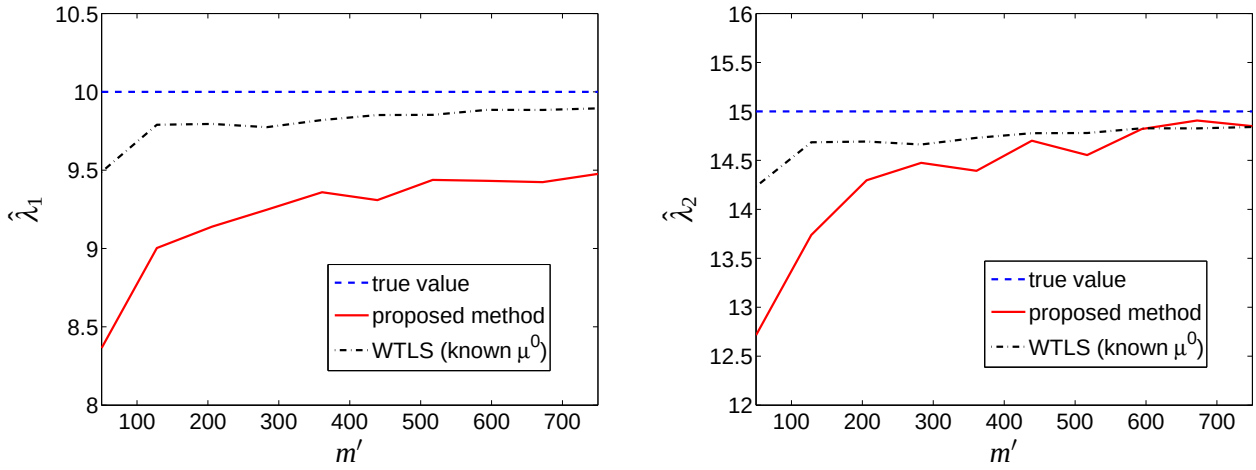

Figure 1: Average values of the noise variance estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ as a function of half the sample size m'. The dashed lines indicate the true values.

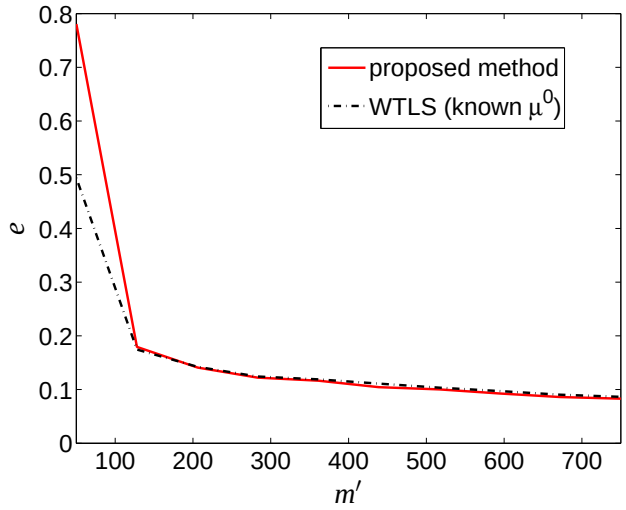

Figure 2: Relative error of estimation e as a function of half the sample size m'.

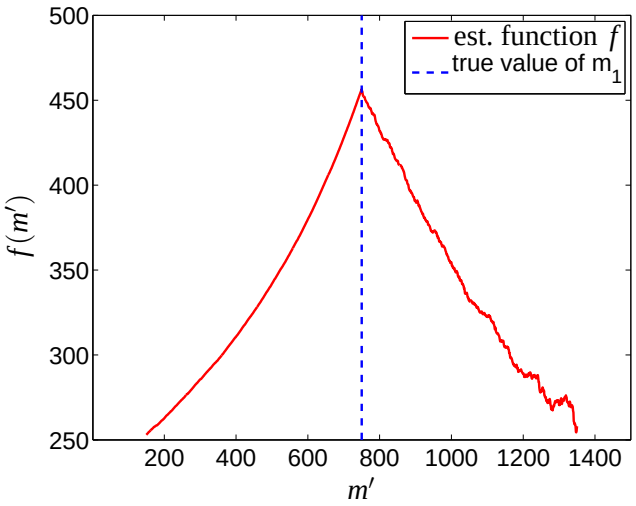

Figure 3: Estimation of the turnover point m_1.

8 Conclusions

We considered a multivariate errors-in-variables model $AX = B$, with finite dependent rows. The total error covariance structure of data matrix $D = \begin{bmatrix} A & B \end{bmatrix}$ was known up to two scalar factors. We supposed that the row data are clustered into two groups, with essentially different second empirical moments. Based on a matrix $D^\top D$, we constructed the consistent estimators of the scalar factors and X.

The clustering assumption is crucial for the consistency. As possible practical application consider a dynamical system where input and output processes are measured with errors. Let both measurement error processes be stationary with rational spectral density and the true input have similar covariance structure. Assume that the input error correlation function is known up to a constant, and the output error correlation function is known up to another constant. Moreover suppose that the input process changes its parameters at certain moment. In such a situation, given the data matrix D we first have to check a statistical hypothesis about the presence of a change point. If the hypothesis is accepted then we have to divide the rows of D in two clusters, according to the optimization problem (8)–(9). After that we can compute the corresponding estimators.

The idea seems plausible for practical system identification. It can be easily generalized to $N > 2$ uncorrelated blocks of the error matrix and N unknown scalars, but then we need N clusters as well, that correspond to $N-1$ change points in the data.

Acknowledgments

Alexander Kukush is a full professor at the Kiev National Taras Shevchenko University, Kiev, Ukraine. He is supported by a senior postdoctoral fellowship from the Department of Applied Economics of the Katholieke Universiteit Leuven. Dr. Ivan Markovsky is a postdoctoral researcher and Dr. Sabine Van Huffel is a full professor at the Katholieke Universiteit Leuven, Belgium.

Our research is supported by Research Council KUL: GOA-AMBioRICS, GOA-Mefisto 666, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0078.01 (structured matrices), G.0407.02 (support vector machines), G.0269.02 (magnetic resonance spectroscopic imaging), G.0270.02 (nonlinear Lp approximation), G.0360.05 (EEG signal processing), research communities (ICCoS, ANMMM); IWT: PhD Grants; Belgian Federal Science Policy Office IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling'); EU: PDT-COIL, BIOPATTERN, ETUMOUR.

The authors are grateful to Prof. H. Schneeweiss for fruitful discussions.

A Proof of Theorem 4.1

Let $m_1 = \kappa m, \kappa \in [r_1, r_2]$. Define a function $\Phi_t(\kappa), \kappa \in [r_1, r_2]$, related to $F(m_1)$:

(a) $\Phi_t(\frac{k}{m}) = F(k)$ if $r_1 \le \frac{k}{m} \le r_2$;

(b) let k_1 and k_2 be the smallest and largest numbers satisfying the latter inequality, then we set $\Phi_t(x) = \Phi_t(\frac{k_1}{m})$, $x \in [r_1, \frac{k_1}{m}]$, and $\Phi_t(x) = \Phi_t(\frac{k_2}{m})$, $x \in [\frac{k_2}{m}, r_2]$;

(c) for $x \in (\frac{i}{m}, \frac{i+1}{m})$, $k_1 \le i \le k_2$, we define $\Phi_t(x)$ by linear interpolation of $\Phi_t(\frac{i}{m})$ and $\Phi_t(\frac{i+1}{m})$.

Now, $\hat{m}_1 = \hat{\kappa} m$, and

$$\Phi_t(\hat{\kappa}) = \max_{r_1 \le \kappa \le r_2} \Phi_t(\kappa). \tag{14}$$

Using (5) and (7), it is possible to show that these exist Ω_0, $\Pr(\Omega_0) = 1$, such that for each $\kappa \in [r_1, r_2]$ and $\omega \in \Omega_0$,

$$\Phi_t(\kappa) \to \Phi_\infty(\kappa) := \varphi(\kappa)\sigma_1(\mathbf{E}\, uu^\top - \mathbf{E}\, vv^\top), \quad \text{as} \quad t \to \infty,$$

where

$$\varphi(\kappa) := \begin{cases} \dfrac{1-r}{1-\kappa}, & \text{if} \quad r_1 \le \kappa \le r \\ \dfrac{r}{\kappa}, & \text{if} \quad r \le \kappa \le r_2. \end{cases}$$

Moreover Ω_0 can be constructed in such a way that for each $\omega \in \Omega_0$ and $\delta > 0$, the functions $\{\Phi_t(\kappa),\ t \geq 1\}$ are equicontinuous on $[r_1, r-\delta] \cup [r+\delta, r_2]$. Thus for $\omega \in \Omega$,

$$\Phi_t(\kappa) \to \Phi_\infty(\kappa), \quad \text{as} \quad t \to \infty, \tag{15}$$

uniformly on $[r_1, r-\delta] \cup [r+\delta, r_2]$, and $\Phi_\infty \in C[r_1, r_2]$. Due to condition (cl_2) the limit function $\Phi_\infty(\kappa)$ has a unique maximum at $\kappa_0 = r$. Then (14) and (15) imply that for each $\omega \in \Omega_0$, $\hat{\kappa} \to \kappa_0$, as $t \to \infty$.Therefore

$$\frac{\hat{m}_1}{m} \to r, \quad \text{as} \quad t \to \infty, \quad \text{a.s.}$$

□

B Proof of Theorem 5.1

B.1 Behavior of $Q_c(\lambda^0)$

We have

$$Q_c(\lambda^0) = \sum_{k=1}^{2} \sum_{i=1}^{p} \mu_{ik}^2(\lambda^0) + c\|\sin\Theta(\lambda^0)\|^2. \tag{16}$$

By Lemma 2.1 and Theorem 4.1, for $k = 1, 2$

$$\left\| \frac{1}{m_k(t)} \Psi_c^{(k)}(\lambda^0) - \frac{1}{m_k(t)} \bar{D}^\top(k) \bar{D}(k) \right\| \to 0, \quad \text{as} \quad t \to \infty, \quad \text{a.s.}$$

We have $\mu_i\big(\bar{D}^\top(k)\bar{D}(k)/m_k(t)\big) = 0$, $1 \leq i \leq p$, and by Lemma 2.2,

$$\lim_{t\to\infty} \mu_{p+1}\left(\frac{1}{m_k(t)} \bar{D}^\top(k)\bar{D}(k) \right) > 0, \quad \text{a.s.}$$

Then a.s.

$$\lim_{t\to\infty} \sum_{k=1}^{2} \sum_{i=1}^{p} \mu_{ik}^2(\lambda^0) = 0,$$

and by Wedin's theorem, see Steward and Sun (1990), Theorem 4.1, p.260,

$$\|\sin\Theta_k(\lambda^0)\| \to 0, \quad \text{as} \quad t \to \infty, \quad \text{a.s.}; \quad k = 1, 2.$$

Here $\Theta_k(\lambda^0)$ is the diagonal matrix of canonical angles between $L_p(\Psi_c^{(k)}(\lambda^0)/m_k(t))$ and $L_p(\bar{D}^\top(k)\bar{D}(k)/m_k(t))$, and L_p denotes the span of the p eigenvectors. Now,

$$L_p(\bar{D}^\top(1)\bar{D}(1)/m_1(t)) = L_p(\bar{D}^\top(2)\bar{D}(2)/m_2(t)) = \text{span}(z_1, \ldots, z_p),$$

where $\begin{bmatrix} z_1 & \ldots & z_p \end{bmatrix} = \begin{bmatrix} X \\ -I_p \end{bmatrix}$. Therefore $\|\sin\Theta(\lambda^0)\| \to 0$, as $t \to \infty$, a.s., and from (16) we obtain

$$Q_c(\lambda^0) \to 0, \quad \text{as} \quad t \to \infty, \quad \text{a.s.}$$

By the definition of $\hat{\lambda}$, this implies that

$$Q_c(\hat{\lambda}) \to 0, \quad \text{as} \quad t \to \infty, \quad \text{a.s.} \tag{17}$$

B.2 $\hat{\lambda}$ is eventually bounded

Now we want to construct such a nonrandom $L > 0$ that eventually

$$\|\hat{\lambda}\| \le L. \tag{18}$$

First from (17) we have for any $\varepsilon_0 > 0$ that

$$\sum_{k=1}^{2}\sum_{i=1}^{p} \mu_{ik}^2(\hat{\lambda}) \le \varepsilon_0 \quad \text{eventually.} \tag{19}$$

We have by Lemma 2.1 and Theorem 4.1 that

$$\left| \frac{1}{m_1(t)} \Psi_c^{(1)}(\hat{\lambda}) - \frac{1}{m_1(t)} \Psi_{\rm ls}^{(1)}(\hat{\lambda}) \right| \to 0, \quad \text{as} \quad t \to \infty, \quad \text{a.s.} \tag{20}$$

Here

$$\Psi_{\rm ls}^{(1)}(\lambda) := \bar{D}^\top(1)\bar{D}(1) - \begin{bmatrix} (\lambda_1 - \lambda_1^0)W_1(1) & 0 \\ 0 & (\lambda_2 - \lambda_2^0)W_2(1) \end{bmatrix}, \quad \lambda := (\lambda_1, \lambda_2) \in [0,\infty) \times [0,\infty).$$

We have for $m_1 = m_1(t)$:

$$\mathbf{tr}\left(X_{\rm ext}^\top (\Psi_{\rm ls}^{(1)}(\hat{\lambda})/m_1) X_{\rm ext} \right) = -\mathbf{tr}\left((\hat{\lambda}_1 - \lambda_1^0) X_1^\top \big(W_1(1)/m_1\big) X_1 + (\hat{\lambda}_2 - \lambda_2^0) X_2^\top \big(W_2(1)/m_1\big) X_2 \right).$$

Suppose that $\hat{\lambda}_1 - \lambda_1^0 > L_0$, where $L_0 > 0$. Then

$$\mathbf{tr}\left(X_{\rm ext}^\top (\Psi_{\rm ls}^{(1)}(\hat{\lambda})/m_1) X_{\rm ext} \right) \le -L_0\, \mathbf{tr}\left(X_1^\top \big(W_1(1)/m_1\big) X_1 \right) + \text{const} \cdot \lambda_2^0.$$

But from (f) and (g),

$$\liminf_{t\to\infty}\ \mathbf{tr}\left(X_1^\top (W_1(1)/m_1) X_1 \right) > 0.$$

Therefore for L_0 large enough and all $t \ge t_0$, we have

$$\mathbf{tr}\left(X_{\rm ext}^\top (\Psi_{\rm ls}^{(1)}(\hat{\lambda})/m_1) X_{\rm ext} \right) \le -1.$$

This implies that $\mu_1\big(\Psi_{\rm ls}^{(1)}(\hat{\lambda})/m_1\big) < 0$ and separated from 0. Then from (20) we have that $\mu_1\big(\Psi_c^{(1)}(\hat{\lambda})/m_1\big)$ is also negative and separated from 0 for $t \ge t_1(\omega)$, a.s. But this contradicts (19). Thus for large enough nonrandom L_0, $\hat{\lambda}_1 - \lambda_1^0 \le L_0$, eventually. Similar inequality can be shown for $\hat{\lambda}_2 - \lambda_2^0$, based on condition (f) for $j = 2$. Thus (18) holds eventually.

B.3 Consistency

We fix Ω_0, $\Pr(\Omega_0) = 1$, such that for all $\omega \in \Omega_0$, $\|\hat{\lambda}(\omega)\| \le L$, for all $t \ge t_0(\omega)$. Fix $\omega \in \Omega_0$. We consider a bounded sequence

$$\{\, \hat{\lambda}(\omega; t) :\ t \ge t_0(\omega) \} \tag{21}$$

and want to show that it converges to λ^0, as $t \to \infty$. Let

$$\{\, \hat{\lambda}\big(\omega; t(q)\big),\ q \ge 1 \}$$

be any convergent subsequence. It means that $t(q) \to \infty$, and

$$\lim_{q\to\infty} \hat{\lambda}\big(\omega; t(q)\big) = \lambda^\infty,$$

for certain $\lambda^\infty \in \mathbb{R}^2$. To prove the desired convergence (21), it suffices to show that $\lambda^\infty = \lambda^0$.

Let

$$M^{(k)}(t) = \text{diag}(\mu_{1k}, \mu_{2k}, \ldots, \mu_{pk})$$

and $Z^{(k)}(t)$ be a matrix $\left[f_1^{(k)}(t) \quad \dots \quad f_p^{(k)}(t)\right]$, of which the columns are the first eigenvectors of $\Psi_c^{(k)}(\hat{\lambda})/m_k(t)$; these columns form an orthogonal basis for $L_{pk}(\hat{\lambda})$. Due to (17), $M^{(k)}(t) \to 0$, as $t \to \infty$, and we can assume that

$$\left\| Z^{(1)}(t) - Z^{(2)}(t) \right\| \to 0, \quad \text{as} \quad t \to \infty. \tag{22}$$

We have

$$\frac{1}{m_k} \Psi_c^{(k)}(\hat{\lambda}) Z^{(k)}(t) = M^{(k)}(t) Z^{(k)}(t). \tag{23}$$

We set here $t = t(q)$. We may and do assume that $Z^{(k)}\big(t(q)\big) \to Z_\infty^{(k)}$, as $q \to \infty$, $k = 1,2$. (Otherwise we should consider a convergent subsequence). From (22) we obtain $Z_\infty^{(1)} = Z_\infty^{(2)} =: Z_\infty$, and from (23) we have, since

$$\sup_{\|\lambda\| \le L} \left\| \frac{1}{m_k(t)} \Psi_c^{(k)}(\lambda) - \frac{1}{m_k(t)} \Psi_{\text{ls}}^{(k)}(\lambda) \right\| \to 0, \quad \text{as} \quad t \to \infty,$$

that

$$\frac{1}{m_k(t(q))} \Psi_{\text{ls}}^{(k)}(\lambda^\infty) Z_\infty \to 0, \quad \text{as} \quad q \to \infty, \quad k = 1,2.$$

Hence

$$\left(\frac{1}{m_1(t(q))} \Psi_{\text{ls}}^{(1)}(\lambda^\infty) - \frac{1}{m_2(t(q))} \Psi_{\text{ls}}^{(2)}(\lambda^\infty) \right) Z_\infty \to 0, \quad \text{as} \quad q \to \infty.$$

Using condition (g), we obtain

$$\left(\frac{1}{m_1(t(q))} \bar{D}^\top(1) \bar{D}(1) - \frac{1}{m_2(t(q))} \bar{D}^\top(2) \bar{D}(2) \right) Z_\infty \to 0, \quad \text{as} \quad q \to \infty.$$

But then from condition (cl_2) similarly to Lemma 2.2 we obtain that $Z_\infty = \left[z_1^\infty \quad \dots \quad z_p^\infty\right]$ with $\operatorname{span}(z_1^\infty, \dots, z_p^\infty) = \operatorname{span}(z_1, \dots, z_p)$. Thus in (23) we have

$$\frac{1}{m_k(t(q))} \Psi_{\text{ls}}^{(k)}(\lambda^\infty) X_{\text{ext}} \to 0, \quad \text{as} \quad q \to \infty, \quad k = 1,2.$$

Thus

$$\begin{bmatrix} (\lambda_1^\infty - \lambda_1^0) W_1(1)/m_1(t(q)) & 0 \\ 0 & (\lambda_2^\infty - \lambda_2^0) W_2(1)/m_2(t(q)) \end{bmatrix} X_{\text{ext}} \to 0$$

and

$$(\lambda_j^\infty - \lambda_j^0) \operatorname{\mathbf{tr}} \left(X_j^\top W_j(1) X_j / m_j(t(q)) \right) \to 0, \quad \text{as} \quad t \to \infty, \quad j = 1,2.$$

But then conditions (f) and (g) imply that $\lambda_j^\infty = \lambda_j^0$, $j = 1,2$. Thus any convergent subsequence of the bounded sequence (21) converges to λ^0, therefore the sequence (21) itself converges to λ^0. This convergence holds for all $\omega \in \Omega_0$, with $\Pr(\Omega_0) = 1$, therefore $\hat{\lambda} \to \lambda^0$, as $t \to \infty$, a.s. □

C Proof of Theorem 6.1

By Theorem 5.1

$$\left\| \frac{1}{m(t)} (\hat{H} - \bar{D}^\top \bar{D}) \right\| \to 0, \quad \text{as} \quad t \to \infty, \quad \text{a.s.}$$

We have

$$\mu_1 \left(\frac{1}{m(t)} \bar{D}^\top \bar{D} \right) = \dots = \mu_p \left(\frac{1}{m(t)} \bar{D}^\top \bar{D} \right) = 0,$$

and by condition (h) and Lemma 2.2, $\mu_{p+1}\left(\bar{D}^\top \bar{D}/m\right)$ is separated from 0, as $t \to \infty$. Moreover, the kernel equals

$$L_p \left(\frac{1}{m} \bar{D}^\top \bar{D} \right) = \operatorname{span}(z_1, \dots, z_p).$$

By Wedin's theorem, see Steward and Sun(1990), Theorem 4.1, p. 260, this implies that $\Theta \to 0$, as $t \to \infty$, where Θ is the diagonal matrix of canonical angles between $L_p(\hat{H})$ and $\text{span}(z_1, \ldots, z_p)$. Moreover $L_p(\hat{H})$ is uniquely defined eventually. Thus $\hat{X}$ is also uniquely defined eventually. Since

$$\begin{bmatrix} X \\ -I_p \end{bmatrix} = \begin{bmatrix} z_1 & \ldots & z_p \end{bmatrix},$$

from $\Theta \to 0$ a.s., we obtain that $\hat{X} \to X$, as $t \to \infty$, a.s. □

References

[1] J. C. Aqũero, and G. C. Goodwin (2006). Idetifiability of errors in variables dynamic systems, *14th IFAC Sympsium on System Identification, Newcastle, Australia.*

[2] B. Anderson (1985). *Identification of scalar errors-in-variables models with dynamics,* Automatica, **21**, 625–755.

[3] R. Carroll, D. Ruppert, and L. Stefanski (1995). *Measurement Error in Nonlinear Models,* Chapman & Hall/CRC, London.

[4] G. Cirrincione, S. Van Huffel, A. Premoli, and M.-L. Rastello (2001). Iteratively reweighted total least squares algorithms for different variances in observations and data, in *Advanced Mathematical & Computational Tools in Metrology V*(eds. P. Ciarlini et al.), 77–84, World Scientific, London.

[5] P. Doukhan (1934). *Mixing. Properties and Examples,* Springer–Verlag, New–York.

[6] A. Edelman, T. A. Arias, and S. T. Smith (1998). The geometry of algorithms with orthogonality constraints, *SIAM J. Matrix Anal. Appl.*, **20**, 303–353.

[7] R. Frisch (1934). *Statistical confluence analysis by means of complete regression systems,* Technical Report 5. Univ. of Oslo, Economics Institute.

[8] A. Kukush and S. Van Huffel (2004). Consistency of elementwise-weighted total least squares estimator in a multivariate errors-in-variables model $AX = B$, *Metrika*, **59**,75–97.

[9] A. Kukush, S. Zwanzig (2001). About the adaptive minimum contrast estimator in a model with nonlinear functional relations, Ukrainian Mathematical Journal, **53**, 1145–1452.

[10] A. Kukush, I. Markovsky, and S. Van Huffel (2005a). Estimation in a linear multivariate measurement error model with clusterin in the regression, Technical Report 05-170, Dept. EE, K.U.Leuven.

[11] A. Kukush, I. Markovsky, and S. Van Huffel (2005b). Consistency of the structured total least squares estimator in a multivariate errors-in-variables model, *J. of Stat. Planning and Inference*, **133**,315–358.

[12] A. Madansky (1959). The fitting of straight lines when both variables are subject to error, Journal of the American Statistical Association, **54**, 173–205.

[13] I. Markovsky, A. Kukush, and S. Van Huffel (2006). On errors-in-variables estimation with unknown noise variance ratio, in the proceedings of the *14th IFAC Symposium on System Identification*, pages 172–177, Newcastle, Australia, 2006.

[14] B. De Moor (1988). Mathematical concepts for modeling of static and dynamic systems. PhD thesis, Dept. EE, K.U.Leuven, Belgium.

[15] J. A. Nelder and R. Mead (1965). A simplex method for function minimization, *Computer J.*, **7**, 308–313.

[16] A. Pakes (1982). On the asymptotic bias of Wald-type estimators of a straight line when both variables are subject to error, International Economic Rewiev, **23**, 491–497.

[17] G. Stewart and J. Sun (1990). *Matrix Perturbation Theory*, Academic Press, Boston.

[18] T. Söderström, U. Soverini, and K. Mahata (2002). Perspectives on errors-in-variables estimation for dynamic systems, *Signal Processing*, **82**, 1139–1154.

[19] A. Wald (1940). The fitting of straight lines if both variables are subject to error, *Annals of Mathematical Statistics*, **11**, 284–300.