Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Statistics & Data Analysis 50 (2006) 181 – 209

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

The element-wise weighted total least-squares problem

Ivan Markovsky[a,*], Maria Luisa Rastello[b], Amedeo Premoli[c], Alexander Kukush[a,1], Sabine Van Huffel[a]

[a] *Dept. Elektrotechneik, ESAT-CSD(SISTA), K.U. Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*
[b] *Istituto Elettrotecnico Nazionale G. Ferraris, Strada delle Cacce 91, I-10135 Torino, Italy*
[c] *Dipartimento di Elettronica e Informazione, Politecnico di Milano Piazza Leonardo da Vinci 32, I-20133 Milano, Italy*

Available online 19 August 2004

Abstract

A new technique is considered for parameter estimation in a linear measurement error model $AX \approx B$, $A = A_0 + \tilde{A}$, $B = B_0 + \tilde{B}$, $A_0X_0 = B_0$ with row-wise independent and non-identically distributed measurement errors $\tilde{A}$, $\tilde{B}$. Here, A_0 and B_0 are the true values of the measurements A and B, and X_0 is the true value of the parameter X. The total least-squares method yields an inconsistent estimate of the parameter in this case. Modified total least-squares problem, called element-wise weighted total least-squares, is formulated so that it provides a consistent estimator, i.e., the estimate $\hat{X}$ converges to the true value X_0 as the number of measurements increases. The new estimator is a solution of an optimization problem with the parameter estimate $\hat{X}$ and the correction $\Delta D = [\Delta A \;\; \Delta B]$, applied to the measured data $D = [A \;\; B]$, as decision variables. An equivalent unconstrained problem is derived by minimizing analytically over the correction ΔD, and an iterative algorithm for its solution, based on the first order optimality condition, is proposed. The algorithm is locally convergent with linear convergence rate. For large sample size the convergence rate tends to quadratic.

Keywords: Total least squares; Multivariate errors-in-variables model; Unequally sized errors; Non-convex optimization; Re-weighted least-squares

[*] Corresponding author. Tel.: +32-16-32-17-10; fax: +32-16-32-19-70.
E-mail address: ivan.markovsky@esat.kuleuven.ac.be (I. Markovsky).
URL: http://www.esat.kuleuven.ac.be/scd.

[1] On leave from National Taras Shevchenko University, Vladimirskaya st. 64, 01601, Kiev, Ukraine.

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.07.014

1. Introduction

Mathematical models are often specified by a set of algebraic, differential, or difference equations. The equations are obtained through a modeling process which is application area dependent. In general, however, the model equations contain unspecified constants that have to be determined from other measurable variables. This process of parameter estimation tunes the model to the measurements (i.e., to the real-life phenomenon) and is of primary interest in many scientific areas.

In this paper, we consider static linear models, i.e., models described by a linear algebraic system of equations $AX = B$. Here $D := [A \;\; B] \in \mathbb{R}^{m\times(n+l)}$ contains the measured data and $X \in \mathbb{R}^{n\times l}$ is the parameter matrix, to be estimated. With less parameters than equations and with noisy data the model equations will not be exactly satisfied, so an approximate solution for X is sought.

The parameter estimation problem is typically defined as an optimization problem: an appropriate cost function depending on the data is minimized over the estimated parameters. The classical approach, the least-squares (LS) estimation technique, minimizes the Frobenius norm of the residual $R = AX - B$. The LS method can be viewed as applying correction ΔB to the right-hand side B in order to make the corrected system $AX = B + \Delta B$ solvable. The correction with the smallest Frobenius norm is sought. Indeed, the LS method is the best linear unbiased estimator when A is noise free and B is corrupted by independent and identically distributed (i.i.d.) errors. We make the assumption that there is a true but unknown value $D_0 = [A_0 \;\; B_0]$ of the measured data and a true value X_0 of the parameter that satisfy the equation $A_0X_0 = B_0$. Moreover, we assume that the measured data D is obtained from the true value with an additive noise $\tilde{D} = [\tilde{A} \;\; \tilde{B}]$, i.e., $D = D_0 + \tilde{D}$. Models of this type are known in the literature (Fuller, 1987; Cheng and Van Ness, 1999) as measurement error (also called errors-in-variables) models.

The total least-squares (TLS) technique (Golub and Van Loan, 1980; Van Huffel and Vandewalle, 1991) is proposed as a parameter estimation technique for the static linear measurement error model when all elements of the data matrix D are perturbed by i.i.d. errors. In the TLS method, a correction $\Delta D = [\Delta A \;\; \Delta B]$ is applied on the matrix D, so that the corrected system of equations $(A + \Delta A)X = B + \Delta B$ becomes solvable. Again the smallest correction, measured by the Frobenius norm, is sought.

The TLS method became popular in the 1980s because the properties of the estimator are well understood, see the monograph Van Huffel and Vandewalle (1991), and robust and efficient methods exist for its solution, based on the singular value decomposition (SVD). The TLS solution $\hat{X}$ is given analytically in terms of the l smallest right singular vectors of the data matrix D. It provides a consistent estimator for the true parameter value X_0 under mild additional assumptions. Consistency means that the estimate $\hat{X}$ converges to X_0 as the number m of the measurements increases.

In the 1990s, a number of extensions of the TLS method have been developed, in order to extend consistency of the TLS estimator to more general noise conditions. Some of the most important contributions are collected in the proceedings (Van Huffel, 1997; Van Huffel and Lemmerling, 2002) of two TLS meetings held in Leuven.

We outline the work connected to the topic of the present paper. In a number of applications, the errors on the elements of D are differently sized. This motivates an extension of the TLS method that relaxes the i.i.d. assumption for the errors. In the so-called generalized total least-squares (GTLS) estimator, the errors $\tilde{D}$ are assumed row-wise independent and correlated within the rows with identical covariance matrix $V = V^\top > 0$. The GTLS problem can be reduced to the TLS problem by post-multiplying the data matrix by $V^{-1/2}$, the inverse of the matrix square root of V (In the actual computation, $V^{1/2}$ is replaced by the computationally cheaper Cholesky factor of V, i.e., the upper triangular matrix U, such that $V = U^\top U$.) This transformation approach, however, is not recommended when the covariance matrix V is ill-conditioned because of the possible loss of accuracy in forming the product $DV^{-1/2}$. In Van Huffel and Vandewalle (1989), a special method is developed, based on the generalized SVD, that makes the scaling implicit and allows a reliable computation of the GTLS estimator.

The GTLS method is still restrictive for some applications because of the assumption that all rows of $\tilde{D}$ have equal covariance matrix. A further generalization for the case when the elements of $\tilde{D}$ are independent, but not identically distributed with element-wise different error variances is proposed in De Moor (1993, Section 4.1). The problem in De Moor (1993, Section 4.1) is univariate (i.e., $l = 1$) and is called element-wise-weighted total least-squares (EW-TLS). In Premoli and Rastello (2002), an algorithm for the computation of the EW-TLS estimator is proposed. The convergence properties of this algorithm, however, are not analyzed. In Kukush and Van Huffel (2004), the EW-TLS problem is generalized to the multivariate case (i.e., $l \geqslant 1$). In addition, the errors are assumed to be row-wise correlated with known covariance matrices V_i, $i = 1, \dots, m$. In the same paper, the multivariate EW-TLS estimator is proven to be statistically consistent.

The formulation of the EW-TLS method is similar to that of the TLS method. Again a correction ΔD that makes the system $(A + \Delta A)X = B + \Delta B$ solvable is introduced, but the cost function is a "weighted Frobenius norm" of the correction. Let $\Delta d_i^\top$ be the ith row of ΔD, i.e., $\Delta D^\top := [\Delta d_1 \ \cdots \ \Delta d_m]$. The EW-TLS cost function is $\sum_{i=1}^m \|V_i^{-1/2} \Delta d_i\|_2^2$. When $V_i = I$, for all i, the EW-TLS cost function reduces to the TLS cost function, and when $V_i = V$, for all i, the EW-TLS cost function reduces to the GTLS cost function.

The EW-TLS estimator generalizes the TLS estimator and improves its statistical accuracy under more general noise assumptions, but makes the problem computationally more difficult. Indeed, while the TLS problem has a closed-form analytical solution and can be computed reliably via the singular value decomposition, the EW-TLS problem has no closed-form solution and its computation involves solving a non-convex optimization problem. For its computation, we propose an iterative algorithm, based on the first-order optimality condition. The convergence depends on the initial approximation. As initial approximation, we propose the GTLS estimator obtained with $V := \sum_{i=1}^m V_i/m$. The GTLS estimator is inconsistent in statistical sense, so an improvement is expected by applying the iterative algorithm starting from this initial approximation.

The contribution of the present paper is a new, more general, formulation of the EW-TLS estimation problem. We allow correlation among the errors within each row of $\tilde{D}$ with possibly singular covariance matrices. (A singularity of the covariance matrix implies error free elements.) We simplify the resulting optimization problem by minimizing analytically over part of the decision variables, those in the correction matrix ΔD. The equivalent problem is an unconstrained optimization problem with less decision variables, namely those in the estimate $\hat{X}$.

Another contribution of the paper is the proposed iterative algorithm for the solution of the equivalent optimization problem. It is a generalization of the algorithm of Premoli and Rastello (2002) for the present more general EW-TLS problem. We prove local convergence with linear convergence rate. For large sample size the convergence rate tends to quadratic. Comparison with standard optimization methods for local optimization (Nelder–Mead simplex method, BFGS quasi–Newton method, and Levenberg–Marquardt method) shows that the proposed algorithm is computationally more efficient for all tested examples.

In order to further motivate the applicability of the presented problem, we show three examples in which the TLS and the GTLS methods are not adequate and a more general problem formulation is called for.

Example 1 (*Relative error TLS*). The correction matrix ΔD is an estimate of $-\tilde{D}$. The TLS cost function $||\Delta D||_F^2 = \sum_{i=1}^m \sum_{j=1}^n \Delta d_{ij}^2$, is a measure of the estimated absolute error ΔD. The relative error TLS problem is defined as: find the minimum correction relative to the given data that makes the system solvable, i.e.,

$$\min_{X,\Delta D} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{\Delta d_{ij}}{d_{ij}} \right)^2 \quad \text{s.t. } (D + \Delta D) \begin{bmatrix} X \\ -I \end{bmatrix} = 0. \tag{1}$$

This problem is an EW-TLS problem with $V_{\tilde{d}_i}^{1/2} := \operatorname{diag}(d_{i1}, \ldots, d_{i(n+l)})$. It is a TLS problem only when $V_{\tilde{d}_i} = \sigma^2 I$, for all i and for certain σ^2, and it is a GTLS problem only when $V_{\tilde{d}_i} = V$, for all i and for certain V.

Example 2 (*Numerical solution of Fredholm integral equations of the first kind*). A Fredholm integral equation of the first kind is

$$\int_{t_a}^{t_b} k_0(s,t) u_0(t)\, \mathrm{d}t = g_0(s) \quad \text{for } s \in [s_a, s_b]. \tag{2}$$

The function g_0 and the kernel k_0 are given and the function u_0 is unknown. Integral equations of the form (2) appear in many scientific and engineering areas, e.g. electrostatics, remote sensing, mathematical biology, and image restoration. An analytic solution is rarely possible, so a numerical approach is typically needed.

In real-life applications, the true data g_0 and k_0 are not exactly known. The function g_0 is measured with additive noise $\tilde{g}$, so given is the noisy counterpart $g = g_0 + \tilde{g}$. The kernel function k_0 is also uncertain with uncertainty modeled by $k = k_0 + \tilde{k}$. In this case, the problem of solving (2) becomes an estimation problem.

Suppose that m measurements are taken for values of s, $\{s_1, \ldots, s_m\} \subset [s_a, s_b]$, and define $\theta_i(t) := k(s_i, t)$, $\tilde{\theta}_i(t) := \tilde{k}(s_i, t)$ for $i = 1, \ldots, m$, where in general the covariance structure of $\tilde{\theta}_i(t)$ depends on i. Suppose also that the solution u is sought in the form of a linear combination of known basis functions $\{f_j\}_{j=1}^n$, i.e., $u(t) = \sum_{j=1}^n x_j f_j(t)$. Then the estimation problem becomes the problem of solving a linear system of equations

$$\begin{bmatrix} \int_{t_\mathrm{a}}^{t_\mathrm{b}} \theta_1(t) f_1(t)\,\mathrm{d}t & \cdots & \int_{t_\mathrm{a}}^{t_\mathrm{b}} \theta_1(t) f_n(t)\,\mathrm{d}t \\ \vdots & & \vdots \\ \int_{t_\mathrm{a}}^{t_\mathrm{b}} \theta_m(t) f_1(t)\,\mathrm{d}t & \cdots & \int_{t_\mathrm{a}}^{t_\mathrm{b}} \theta_m(t) f_n(t)\,\mathrm{d}t \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \approx \begin{bmatrix} g(s_1) \\ \vdots \\ g(s_m) \end{bmatrix} \tag{3}$$

with additive errors in both right-hand side and coefficient matrix. If the errors $\tilde{\theta}_i$ and $\tilde{g}(s_i)$ are independent for different i, i.e., the errors are independent from measurement to measurement, then solving (3) in the maximum-likelihood sense is an EW-TLS problem. Clearly, the TLS and GTLS estimates are maximum-likelihood ones only in special cases. They can be used, however, in the general case, to find suboptimal solutions.

Example 3 (*Application in mineralogy*). Another realistic example appears in mineralogy, see Fisher (1989). Fisher applies the TLS technique for analysis of metamorphic assemblages. He uses diagonal scaling to take into account differently sized errors but, as quoted below (Fisher, 1989, p. 74), he recognized that a more general method is needed.

> Though simple to apply, this technique for weighting the composition matrix is not ideal; only rarely will the matrix of estimated uncertainties have the structure of a product of two diagonal matrices. Further research into techniques of weighting the composition matrix seems desirable.

The paper is organized as follows. In Section 2, a notation for set indexing used throughout the paper is introduced. In Section 3, the EW-TLS problem is defined. It is an optimization problem with decision variables, the parameter estimate and the correction. In Section 4, we eliminate the correction and derive an equivalent unconstrained optimization problem. The latter is considered in Section 5. An iterative algorithm is proposed based on the first-order optimality condition. The gradient of the cost function is derived in Appendix A. We state and prove local convergence results. The proofs are given in Appendix B and C. In Section 6, we present simulation examples that illustrate the consistency of the EW-TLS estimator and the relative error TLS problem of Example 1. Conclusions are given in Section 7.

2. Notation for set indexing

For a set $\mathscr{S}$, a subset of $\mathscr{U}$, $\bar{\mathscr{S}}$ is the complement of $\mathscr{S}$ relative to $\mathscr{U}$. The universal set $\mathscr{U}$ will be understood from the context.

Given a set of indices $\mathscr{I} \subseteq \{1, \ldots, m\}$ and a vector $a \in \mathbb{R}^m$, $a(\mathscr{I})$ (alternatively $a_{\mathscr{I}}$) denotes the vector derived from a by deleting the elements with indices in $\bar{\mathscr{I}}$. Let $i_1, \ldots, i_k$ be the ordered elements of the set $\mathscr{I}$,

$$\dim \mathscr{I} = k, \quad \mathscr{I} = \{i_1, \ldots, i_k\}, \quad i_1 < i_2 < \cdots < i_k.$$

Define the matrix of unit vectors

$$T(\mathscr{I}) := [\mathbf{1}_{i_1} \cdots \mathbf{1}_{i_k}],$$

where $\mathbf{1}_i \in \mathbb{R}^m$ denotes the ith unit vector. We have,

$$a_{\mathscr{I}} = a(\mathscr{I}) = T(\mathscr{I})^\top a.$$

Similarly, given a pair of sets $\mathscr{I} \subseteq \{1, \ldots, m\}$ and $\mathscr{J} \subseteq \{1, \ldots, n\}$ and a matrix $A \in \mathbb{R}^{m \times n}$, $A(\mathscr{I}, \mathscr{J})$ (alternatively $A_{\mathscr{I}\mathscr{J}}$) denotes the matrix formed from A by deleting the rows with indices in the set $\bar{\mathscr{I}}$ and the columns with indices in the set $\bar{\mathscr{J}}$. Let $\mathscr{I} =: \{i_1, \ldots, i_k\}$, $i_1 < i_2 < \cdots < i_k$, and $\mathscr{J} =: \{j_1, \ldots, j_l\}$, $j_1 < j_2 < \cdots < j_l$. Then

$$A_{\mathscr{I}\mathscr{J}} = A(\mathscr{I}, \mathscr{J}) = T(\mathscr{I})^\top A T(\mathscr{J}).$$

A colon (:) is used instead of either $\mathscr{I}$ or $\mathscr{J}$ to denote, respectively, the set of row indices $\{1, \ldots, m\}$ or the set of column indices $\{1, \ldots, n\}$. For example, $A(i, :)$ is the ith row of A and $A(:, j)$ is the jth column of A.

The transposed ith row of A, $(A(i, :))^\top$, is denoted by a_i, so that we have $A^\top = [a_1 \ \cdots \ a_m]$. The following conventions and rule are used for interchanging set indexing and transposition

$$(A(\mathscr{I}, \mathscr{J}))^\top := A(\mathscr{I}, \mathscr{J})^\top = A(\mathscr{J}, \mathscr{I})^\top =: (A^\top)_{\mathscr{J}\mathscr{I}}.$$

3. Problem formulation

Consider the linear measurement error model

$$AX \approx B, \quad A = A_0 + \tilde{A}, \quad B = B_0 + \tilde{B} \quad \text{with } A \in \mathbb{R}^{m \times n}, \ \ B \in \mathbb{R}^{m \times l}. \tag{4}$$

The matrices A and B are measurements of the true but unobservable A_0 and B_0, and X is a parameter of interest. $\tilde{A}$ and $\tilde{B}$ are measurement errors, respectively. We suppose that there exists a matrix $X_0 \in \mathbb{R}^{n \times l}$, the true value of the parameter, that satisfies (4) exactly,

$$A_0 X_0 = B_0.$$

The measurement errors $\tilde{A}$ and $\tilde{B}$ are random matrices, such that $\tilde{D} := [\tilde{A} \ \tilde{B}]$ has zero mean and independent rows $\tilde{d}_i$ with known row covariance matrices

$$V_{\tilde{d}_i} := \operatorname{cov}(\tilde{d}_i) \quad \text{for } i = 1, \ldots, m.$$

An alternative formulation for the model (4), which we use later, is

$$DX_{\text{ext}} \approx 0, \quad D = D_0 + \tilde{D} \quad \text{with } D \in \mathbb{R}^{m \times (n+l)}, \ \ X_{\text{ext}} \in \mathbb{R}^{(n+l) \times l}. \tag{5}$$

Here $D := [A \ B]$ contains the measured data, $D_0 := [A_0 \ B_0]$ the true data, and $X_{\text{ext}} := \left[\begin{smallmatrix} X \\ -I \end{smallmatrix}\right]$ is the extended parameter matrix. The true value $X_{0,\text{ext}} := \left[\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right]$ of the extended parameter satisfies (5) exactly,

$$D_0 X_{0,\text{ext}} = 0.$$

Given the available measured data D and the error covariance information $\{V_{\tilde{d}_i}\}_{i=1}^m$, corresponding to each row, we aim to estimate the true value X_0 of the parameter and the true data D_0. First, we define the EW-TLS problem assuming that all matrices $V_{\tilde{d}_i}$ are non-singular, which implies that all elements of D are noisy:

$$\min_{X,\,\Delta D} \sum_{i=1}^{m} \| V_{\tilde{d}_i}^{-\frac{1}{2}} \Delta d_i \|_2^2 \quad \text{s.t. } (D + \Delta D) \begin{bmatrix} X \\ -I \end{bmatrix} = 0. \tag{6}$$

Here, ΔD is a correction on the measured data introduced to compensate for the measurement error $\tilde{D}$. The optimization variables are X and ΔD. Let $(\hat{X}, \Delta\hat{D})$ be an optimal point of the EW-TLS problem (6). $\hat{X}$ is an EW-TLS estimate of the true value X_0 of the parameter and $D + \Delta\hat{D}$ is an EW-TLS estimate of the true data D_0.

The proposed estimation method is the maximum-likelihood estimator for the defined model and under mild additional assumptions is statistically consistent, see Kukush and Van Huffel (2004).

Remark 4 (*Covariance known up to a constant*). In the EW-TLS estimation setup, the exact covariances $\{V_{\tilde{d}_i}\}_{i=1}^m$ are not needed; knowledge of $V_{\tilde{d}_i}$ up to a constant factor suffices. Suppose that instead of the covariance matrices $V_{\tilde{d}_i}$, matrices $\{W_{\tilde{d}_i}\}_{i=1}^m$ are given such that $V_{\tilde{d}_i} = \gamma_0 W_{\tilde{d}_i}$ for $i = 1, \ldots, m$ and for some unknown constant γ_0. Then $W_{\tilde{d}_i}$ can be used in place of $V_{\tilde{d}_i}$ in what follows. The cost function of (6) is proportional to $1/\gamma_0$ and the minimum point is not affected.

Remark 5 (*TLS as a special case of the EW-TLS*). For the case where all $d_i, i = 1, \ldots, m$, are perturbed with errors $\tilde{d}_i$ with unit covariance matrix $W_{\tilde{d}_i} = I_{n+l}$, $i = 1, \ldots, m$, known up to a factor of proportionality γ_0, the EW-TLS problem (6) reduces to the TLS problem, i.e.,

$$\min_{X,\Delta D} \sum_{i=1}^{m} ||\Delta d_i||_2^2 = \min_{X,\Delta D} ||\Delta D||_F^2 \quad \text{s.t. } (D + \Delta D) \begin{bmatrix} X \\ -I \end{bmatrix} = 0. \tag{7}$$

Next, we consider a more general EW-TLS problem formulation where some of the elements of D are allowed to be noise free. In this case, some covariance matrices $V_{\tilde{d}_i}$ are singular. Let $\mathscr{J}_i \subseteq \{1, \ldots, n+l\}$ be a set of column indices such that $D(i, \mathscr{J}_i)$ is measured with noise and $D(i, \bar{\mathscr{J}}_i)$, is noise free,

$$\operatorname{var}(\tilde{d}_{ij}) \begin{cases} \neq 0 & \text{if } j \in \mathscr{J}_i \\ = 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, \ldots, m.$$

If $\bar{\mathscr{J}}_i$ is non-empty, then $V_{\tilde{d}_i}(\bar{\mathscr{J}}_i, \bar{\mathscr{J}}_i) = 0$ and $V_{\tilde{d}_i}$ is singular. We define V_i as the covariance matrix of the non-deterministic part of $\tilde{d}_i$,

$$V_i := \operatorname{cov}(\tilde{d}_i(\mathscr{J}_i)) = V_{\tilde{d}_i}(\mathscr{J}_i, \mathscr{J}_i).$$

For $i = 1, \ldots, m$, V_i has full rank. The noise free measurements in the ith row $D(i, \bar{\mathscr{J}}_i)$ do not need any correction, implying that $\Delta D(i, \bar{\mathscr{J}}_i) = 0$. We introduce new variables

$c_i \in \mathbb{R}^{\dim \mathscr{I}_i}$, for the nonzero sub-vector of the corresponding corrections Δd_i,

$$c_i := \Delta d_i(\mathscr{I}_i) = T(\mathscr{I}_i)^\top \Delta d_i \quad \text{for } i = 1, \ldots, m. \tag{8}$$

The EW-TLS problem in the presence of a mixture of noise free and noisy measurements is defined as

$$\min_{\substack{X\\ c_1,\ldots,c_m}} \sum_{i=1}^{m} \|V_i^{-\frac{1}{2}} c_i\|_2^2 \quad \text{s.t. } DX_{\text{ext}} + \begin{bmatrix} c_1^\top X_{\text{ext}}(\mathscr{I}_1, :) \\ \vdots \\ c_m^\top X_{\text{ext}}(\mathscr{I}_m, :) \end{bmatrix} = 0, \quad X_{\text{ext}} = \begin{bmatrix} X \\ -I \end{bmatrix}. \tag{9}$$

Remark 6 (*Noise-free rows*). The presence of noise-free rows in the data matrix D can be used in a pre-processing step in order to reduce the size of the estimation problem. Suppose that k rows of D are noise free. If $k \geqslant n$, the estimation problem becomes trivial, so we suppose in addition that $k < n$. Rearranging the rows of D, so that the last k rows are noise free, we have $D = [\begin{smallmatrix} D_1 \\ D_2 \end{smallmatrix}]$, with $D_2 \in \mathbb{R}^{k \times (n+l)}$, noise free. While $c_i = 0$ for $i = m-k+1, \ldots, m$, (9) can be written as

$$\begin{aligned} &\min_{\substack{X\\ c_1,\ldots,c_{m-k}}} \quad \sum_{i=1}^{m-k} \|V_i^{-\frac{1}{2}} c_i\|_2^2 \\ &\text{s.t.} \quad D_2 X_{\text{ext}} = 0, \quad D_1 X_{\text{ext}} + \begin{bmatrix} c_1^\top X_{\text{ext}}(\mathscr{I}_1, :) \\ \vdots \\ c_{m-k}^\top X_{\text{ext}}(\mathscr{I}_{m-k}, :) \end{bmatrix} = 0, \, X_{\text{ext}} = \begin{bmatrix} X \\ -I \end{bmatrix}. \end{aligned} \tag{10}$$

Let $D_2 := [A_2 \;\; B_2]$. The constraint $A_2 X = B_2$ is equivalent to $X = N\bar{X} + X_{\text{p}}$ for some $\bar{X} \in \mathbb{R}^{(n-k)\times l}$, where N is a matrix of which the columns form a basis for the null space of A_2 and X_{p} is a particular solution of $A_2 X = B_2$. Substituting $N\bar{X} + X_{\text{p}}$ for X in (10) and considering $\bar{X}$ as a new variable, we obtain an equivalent EW-TLS problem without noise-free rows

$$\begin{aligned} &\min_{\substack{\bar{X}\\ c_1,\ldots,c_{m-k}}} \quad \sum_{i=1}^{m-k} \|V_i^{-\frac{1}{2}} c_i\|_2^2 \quad \text{s.t. } D_1 X_{\text{ext}} + \begin{bmatrix} c_1^\top X_{\text{ext}}(\mathscr{I}_1, :) \\ \vdots \\ c_{m-k}^\top X_{\text{ext}}(\mathscr{I}_{m-k}, :) \end{bmatrix} = 0, \\ &X_{\text{ext}} = \begin{bmatrix} N\bar{X} + X_{\text{p}} \\ -I \end{bmatrix}. \end{aligned}$$

The new problem is of smaller dimension, both in terms of number of constraints and number of variables.

4. Minimization with respect to the correction

The first stage in solving the EW-TLS problem is to minimize analytically the cost function with respect to the correction $\{c_i\}_{i=1}^m$, i.e., we find a function $f_0 : \mathbb{R}^{n\times l} \to \mathbb{R}$,

$$f_0(X) := \min_{c_1,\ldots,c_m} \sum_{i=1}^{m} \|V_i^{-\frac{1}{2}} c_i\|_2^2 \quad \text{s.t.} \quad DX_{\text{ext}} + \begin{bmatrix} c_1^\top X_{\text{ext}}(\mathscr{J}_1,:) \\ \vdots \\ c_m^\top X_{\text{ext}}(\mathscr{J}_m,:) \end{bmatrix} = 0, \quad X_{\text{ext}} = \begin{bmatrix} X \\ -I \end{bmatrix} \tag{11}$$

for all $X \in \mathbb{R}^{n\times l}$. As a result the EW-TLS problem (9) becomes the unconstrained optimization problem

$$\min_X \ f_0(X). \tag{12}$$

For a fixed $X \in \mathbb{R}^{n\times l}$, X_{ext} is a fixed given matrix and the constraint of (11) is a linear equation in the optimization variables $\{c_i\}_{i=1}^m$. It can be represented as a set of linear equations in $\{c_i\}_{i=1}^m$.

$$\begin{aligned} DX_{\text{ext}} + \begin{bmatrix} c_1^\top X_{\text{ext}}(\mathscr{J}_1,:) \\ \vdots \\ c_m^\top X_{\text{ext}}(\mathscr{J}_m,:) \end{bmatrix} = 0 \quad &\Leftrightarrow \quad c_i^\top X_{\text{ext}}(\mathscr{J}_i,:) = -(DX_{\text{ext}})_{i,:}, \quad i=1,\ldots,m \\ &\Leftrightarrow \quad X_{\text{ext}}(:,\mathscr{J}_i)^\top c_i = -((DX_{\text{ext}})_{i,:})^\top. \end{aligned}$$

Define

$$G_i(X) := X_{\text{ext}}(:,\mathscr{J}_i)^\top = X_{\text{ext}}^\top T(\mathscr{J}_i)$$

and the residual matrix

$$R(X) := DX_{\text{ext}} = AX - B.$$

Denote by $r_i^\top(X)$ the ith row of $R(X)$, i.e.,

$$R^\top(X) = [r_1(X) \cdots r_m(X)].$$

With this notation, the constraint of (11) is equivalent to

$$G_i(X)c_i = -r_i(X), \quad i=1,\ldots,m,$$

which shows that the optimization problem (11) is separable in c_i. As a consequence, we have to solve m-independent optimization problems

$$f_i(X) = \min_{c_i} ||V_i^{-\frac{1}{2}} c_i||_2^2 \quad \text{s.t. } G_i(X)c_i = -r_i(X), \quad i=1,\ldots,m.$$

The solution of (11) is given by $f_0(X) = \sum_{i=1}^m f_i(X)$. The common problem

$$\min_c \ ||V^{-\frac{1}{2}} c||_2^2 \quad \text{s.t. } G(X)c = -r(X) \tag{13}$$

is a least-norm problem, so that its solution is

$$c_{\mathrm{opt}}(X) = -VG^{\top}(X)(G(X)VG^{\top}(X))^{-1}r(X)$$

and the optimal value is

$$c_{\mathrm{opt}}^{\top}(X)V^{-1}c_{\mathrm{opt}}(X) = r^{\top}(X)(G(X)VG^{\top}(X))^{-1}r(X).$$

Then the solution of (11) becomes

$$f_0(X) = \sum_{i=1}^{m} r_i^{\top}(X)(G_i(X)V_iG_i^{\top}(X))^{-1}r_i(X)$$

(this function is well known, see, e.g., Sprent, 1966) and the optimal correction is

$$c_{i,\mathrm{opt}} = -V_iG_i^{\top}(X)(G_i(X)V_iG_i^{\top}(X))^{-1}r_i(X), \quad i = 1, \ldots, m.$$

While $\tilde{d}_i(\bar{\mathscr{J}}_i)$ is deterministic, $T(\bar{\mathscr{J}}_i)^{\top}V_{\tilde{d}_i}T(\bar{\mathscr{J}}_i) = 0$ and

$$T(\mathscr{J}_i)V_iT(\mathscr{J}_i)^{\top} = V_{\tilde{d}_i}.$$

Using this fact and (8) the solution can be written as

$$f_0(X) = \sum_{i=1}^{m} r_i^{\top}(X)(X_{\mathrm{ext}}^{\top}V_{\tilde{d}_i}X_{\mathrm{ext}})^{-1}r_i(X) \tag{14}$$

and

$$\Delta D_{\mathrm{opt}} = -\begin{bmatrix} r_1^{\top}(X)(X_{\mathrm{ext}}^{\top}V_{\tilde{d}_1}X_{\mathrm{ext}})^{-1}X_{\mathrm{ext}}^{\top}V_{\tilde{d}_1} \\ \vdots \\ r_m^{\top}(X)(X_{\mathrm{ext}}^{\top}V_{\tilde{d}_m}X_{\mathrm{ext}})^{-1}X_{\mathrm{ext}}^{\top}V_{\tilde{d}_m} \end{bmatrix}. \tag{15}$$

Remark 7. The weighting matrices in (14) are the covariance matrices of the residuals, i.e.,

$$X_{\mathrm{ext}}^{\top}V_{\tilde{d}_i}X_{\mathrm{ext}} = \mathrm{var}(r_i(X)) \quad \text{for } i = 1, \ldots, m.$$

Remark 8. The sets $\mathscr{J}_i$, $i = 1, \ldots, m$ do not participate in the solution (14). The optimization problem (12) automatically “recognizes” the noise-free elements in D on the basis of the covariance information $\{V_{\tilde{d}_i}\}_{i=1}^{m}$. The solution of the purely noisy formulation (6), is given again by (14), which shows that the reformulation to the more general error-free case (9) is only needed to avoid the problem of inversion of singular matrices in the derivation.

Remark 9 (*Correction elimination in the TLS case*). Restricting the solution (14), to the TLS case, $W_{\tilde{d}_i} = I$ for all i, we have that (7) is equivalent to

$$\min_{X} \; f_{\mathrm{TLS}}(X), \tag{16}$$

where

$$\begin{aligned} f_{\mathrm{TLS}}(X) &:= \sum_{i=1}^{m} r_i^\top(X)(X_{\mathrm{ext}}^\top X_{\mathrm{ext}})^{-1} r_i(X) \\ &= \operatorname{trace}(R(X)(X_{\mathrm{ext}}^\top X_{\mathrm{ext}})^{-1} R^\top(X)) \end{aligned}$$

and the optimal correction is

$$\Delta D_{\mathrm{TLS}} = -\begin{bmatrix} r_1^\top(X)(X_{\mathrm{ext}}^\top X_{\mathrm{ext}})^{-1} X_{\mathrm{ext}}^\top \\ \vdots \\ r_m^\top(X)(X_{\mathrm{ext}}^\top X_{\mathrm{ext}})^{-1} X_{\mathrm{ext}}^\top \end{bmatrix} = -R(X)(X_{\mathrm{ext}}^\top X_{\mathrm{ext}})^{-1} X_{\mathrm{ext}}^\top.$$

Remark 10 (*Non-convexity of the EW-TLS problem*). The EW-TLS cost function f_0 is non-convex. A simple counter example is the function $f_0(x) = (x-1)^2/(1+x^2)$, which is a special case of (14). Due to the non-convexity of the problem, we consider iterative methods for local optimization.

5. Iterative algorithm

In the rest of the paper we consider the resulting optimization problem (12). For the special case of the TLS problem, (12) becomes (16) and can be solved analytically in terms of the SVD of the data matrix $[A \;\; B]$. In the more general case, however, there is no analytic solution and we rely on a numerical solution method. In Section 5.1, we derive an iterative algorithm. It is based on an approximation of the first-order optimality condition of (12). In Section 5.2, we outline the algorithm and derive a special version for the case when all errors are uncorrelated. In Section 5.3, we state the local convergence results.

5.1. Derivation of the algorithm

The first-order optimality condition of (12) is

$$f_0'(X) = 0. \tag{17}$$

The derivative of f_0 with respect to X is (see Appendix A)

$$\begin{aligned} f_0'(X) = 2\sum_{i=1}^{m} \Bigg(& a_i r_i^\top(X) Q_i^{-1}(X) - \begin{bmatrix} V_{\tilde a_i} & V_{\tilde a_i \tilde b_i} \end{bmatrix} \\ & \times \begin{bmatrix} X \\ -I \end{bmatrix} Q_i^{-1}(X) r_i(X) r_i^\top(X) Q_i^{-1}(X) \Bigg), \end{aligned}$$

where for convenience we set

$$Q_i(X) := X_{\text{ext}}^{\top} V_{\tilde{d}_i} X_{\text{ext}}$$

and the covariance matrix $V_{\tilde{d}_i}$ is partitioned as

$$V_{\tilde{d}_i} = \begin{bmatrix} \operatorname{var}(\tilde{a}_i) & \operatorname{cov}(\tilde{a}_i, \tilde{b}_i) \\ \operatorname{cov}(\tilde{b}_i, \tilde{a}_i) & \operatorname{var}(\tilde{b}_i) \end{bmatrix} =: \begin{bmatrix} V_{\tilde{a}_i} & V_{\tilde{a}_i \tilde{b}_i} \\ V_{\tilde{b}_i \tilde{a}_i} & V_{\tilde{b}_i} \end{bmatrix}.$$

Eq. (17) is a necessary condition for a minimum of (12), i.e., a solution of (17) corresponds to the desired global minimum of (12). Solving (17), however, is a difficult nonlinear problem. The idea we use is to approach a solution of (17) by applying an iterative procedure. Let $X^{(k)}$ be the approximation on the kth step. The approximation $X^{(k+1)}$ on the next step is defined as the solution of the equation

$$F(X^{(k+1)}, X^{(k)}) = 0. \tag{18}$$

Here F is an approximation of $f_0'(X^{(k+1)})$, obtained by fixing X to $X^{(k)}$, in some places where X appears in (17). The choice where to fix X is motivated by the desire to obtain an easier to solve equation. A choice that leads to a linear equation is

$$\begin{aligned} F(X^{(k+1)}, X^{(k)}) \\ := 2 \sum_{i=1}^{m} \Big(& a_i (X^{(k+1)T} a_i - b_i)^{\top} Q_i^{-1}(X^{(k)}) \\ & -(V_{\tilde{a}_i} X^{(k+1)} - V_{\tilde{a}_i \tilde{b}_i}) Q_i^{-1}(X^{(k)}) r_i(X^{(k)}) r_i^{\top}(X^{(k)}) Q_i^{-1}(X^{(k)}) \Big). \end{aligned} \tag{19}$$

The approximation (19) is proposed in Premoli and Rastello (2002), for the special case $l = 1$ and $V_{\tilde{d}_i} = \operatorname{diag}(\sigma_{i1}, \ldots, \sigma_{i(n+1)})$, for all i.

On the kth step of the iterative algorithm, we solve Eq. (18). The process is repeated until $||X^{(k+1)} - X^{(k)}||_F / ||X^{(k+1)}||_F < \varepsilon$, i.e., until the norm of the relative difference between the new estimate and the previous one is smaller than a given tolerance ε.

The algorithm is a successive approximation type algorithm. It is heuristic because Eq. (17) is only a necessary condition for optimality of (12), and the iterative procedure is not guaranteed to converge globally to a solution of (17). We prove, however, local convergence of the iterative procedure and compare its performance numerically with this of standard optimization methods.

Remark 11. The proposed algorithm is *not* a Gauss–Newton-type algorithm for solving Eq. (17) because the proposed approximation F is not the first-order truncated Taylor series of f_0'; it is another linear approximation. Our choice makes the derivation of the algorithm simpler but it turns out that the convergence analysis is more difficult than that for the Gauss–Newton algorithm.

5.2. Algorithm

We give an outline of the algorithm described in Section 5.1.

Algorithm 1 Computation of the EW-TLS estimator $\hat{X}_{\text{EW-TLS}}$.

```
Step 1:  Given the measurements $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times l}$; the error covariance information
         $V_{\tilde{d}_i} \in \mathbb{R}^{(n+l)\times(n+l)}$, $i = 1, \dots, m$; and a convergence tolerance $\varepsilon$.
Step 2:  Find an initial approximation $X^{(0)}$.
Step 3:  Initialize the iteration counter $k := 0$.
Step 4:  repeat
Step 5:    Let $Q_i(X^{(k)}) := [\begin{smallmatrix} x^{(k)} \\ -I \end{smallmatrix}]^\top V_i [\begin{smallmatrix} X^{(k)} \\ -I \end{smallmatrix}]$, for $i = 1, \dots, m$.
Step 6:    Let $R(X^{(k)}) := AX^{(k)} - B$.
Step 7:    Solve the linear system $F(X^{(k+1)}, X^{(k)}) = 0$ for $X^{(k+1)}$.
Step 8:    Increment the iteration counter $k := k + 1$.
Step 9:  until $||X^{(k)} - X^{(k-1)}||/||X^{(k)}|| < \varepsilon$
Step 10: The computed EW-TLS estimator is $\hat{X}_{\text{EW-TLS}} := X^{(k)}$.
```

The computations on Steps 2 and 7 are specified next. The GTLS estimate with weighting matrix $V := \sum_{i=1}^{m} V_i/m$ can be used for the initial approximation $X^{(0)}$. This involves the generalized singular value decomposition of D, see Van Huffel and Vandewalle (1991). Alternatively, the computationally cheaper weighted least-squares (WLS) estimate can be used. The choice depends on the noise covariance information: if the size of the errors in B is relatively larger than the size of the errors in A, then the WLS estimate outperforms the GTLS estimate and should be used as initial approximation.

The computation in Step 7 is the kernel of the algorithm. We consider separately two cases: univariate problems, i.e., $l = 1$, and multivariate problems, i.e., $l > 1$. In the univariate case, $Q_i(x^{(k)})$, $r_i(x^{(k)})$, and b_i are scalars and Step 7 is reduced as follows. Solve

$$\sum_{i=1}^{m} \Big(a_i (a_i^\top x^{(k+1)} - b_i) Q_i^{-1}(x^{(k)}) \\ -(V_{\tilde{a}_i} x^{(k+1)} - V_{\tilde{a}_i \tilde{b}_i}) Q_i^{-1}(x^{(k)}) r_i(x^{(k)}) r_i(x^{(k)}) Q_i^{-1}(x^{(k)}) \Big) = 0,$$

which is equivalent to a standard linear system of equations $G(x^{(k)})x^{(k+1)} = h(x^{(k)})$,

$$\underbrace{\sum_{i=1}^{m} \left(\frac{a_i a_i^\top}{Q_i(x^{(k)})} - V_{\tilde{a}_i} \frac{r_i^2(x^{(k)})}{Q_i^2(x^{(k)})} \right)}_{G(x^{(k)})} x^{k+1} = \underbrace{\sum_{i=1}^{m} \left(\frac{a_i b_i}{Q_i(x^{(k)})} - V_{\tilde{a}_i \tilde{b}_i} \frac{r_i^2(x^{(k)})}{Q_i^2(x^{(k)})} \right)}_{h(x^{(k)})}. \tag{20}$$

In the multivariate case, the equation $F(X^{(k+1)}, X^{(k)}) = 0$ is vectorized and then solved as a standard linear system of equations $G(X^{(k)})\text{vec}(X^{(k+1)}) = h(X^{(k)})$.

If the elements of $\tilde{D}$ are uncorrelated, i.e., $V_{\tilde{d}_i} = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{i(n+1)}^2)$ for all i, a more compact description of the error covariance information is the matrix $\Sigma = [\sigma_{ij}] \in \mathbb{R}^{m\times(n+1)}$ of the element-wise standard deviations. In this case, Step 5 simplifies to $Q_i(x) := \sum_{j=1}^{n} \sigma_{ij}^2 x_j^2 + \sigma_{i(n+1)}^2$.

5.3. *Local convergence*

We list the assumptions used in the theorems. $\lambda_{\max}(V)$ denotes the maximum eigenvalue and $\lambda_{\min}(V)$ denotes the minimum eigenvalue of the symmetric matrix V.

(i) The random vectors $\tilde{d}_i$, $i = 1, \ldots, m$, are independent with zero mean and finite second moments.
(ii) $\mathscr{J}_i = \mathscr{J}$, for $i = 1, \ldots, m$, i.e., if the data matrix D has noise-free elements, then they appear only in noise-free columns.
(iii) There exists a number $\kappa > 0$, such that $\lambda_{\min}(V_i) \geqslant \kappa$, for $i = 1, 2, \ldots, m$.
(iv) $\text{rank}(X_0(\mathscr{J}, :)) = l$.
(v) Let $s := \dim \mathscr{J} - (l+1)/2$. For a fixed real number δ, $\delta \geqslant 2$, $\delta > s$,

$$\sup_{(i\geqslant 1,\ j\in\mathscr{J})} \mathbf{E}|\tilde{d}_{ij}|^{2\delta} < \infty.$$

(vi) $\lambda_{\min}(A_0^\top A_0)/\sqrt{m} \to \infty$, as $m \to \infty$.
(vii) $\lambda_{\min}^2(A_0^\top A_0)/\lambda_{\max}(A_0^\top A_0) \to \infty$, as $m \to \infty$.
(viii) For a fixed real number δ, $\delta \geqslant 2$, $\delta > s$, $\delta > nl$, $\sup_{(i\geqslant 1,\ j\in\mathscr{J})} \mathbf{E}|\tilde{d}_{ij}|^{2\delta} < \infty$.
(ix) $\limsup_{m\to\infty} \lambda_{\max}(A_0^\top A_0)/\lambda_{\min}(A_0^\top A_0) < \infty$.

Assumptions (vi) and (vii) are Gallo's conditions for statistical consistency in the univariate model, see Gallo (1982). Condition (vi) means that $\lambda_{\min}(A_0^\top A_0)$ tends to infinity fast enough. Thus $A_0^\top A_0$ is away from singularity. In particular A_0 is a full rank matrix, for large enough m. Assumption (vii) can be interpreted as follows. Although $\lambda_{\max}(A_0^\top A_0)$ tends to infinity together with $\lambda_{\min}$, it does not tend to infinity "too fast". For example, both conditions hold under the following stability condition

$$A_m = A_0^\top A_0/m \to A_\infty > 0 \quad \text{as } m \to \infty,$$

or in a slightly more general setting, if

$$\liminf_{m\to\infty} \lambda_{\min}(A_m) > 0 \quad \text{and} \quad \limsup_{m\to\infty} \lambda_{\max}(A_m) < \infty.$$

The stability condition means that A_m either converges to a non-singular matrix or it stays away from singularity and is bounded. For example, A_m stabilizes when the rows of A_0 are randomly chosen from a distribution with a non-singular covariance matrix. This example could be regarded as typical for the applications considered in the paper.

Since the original problem is defined for random data, the convergence statement is also in probabilistic terms. Hereafter, $\mathrm{O_p}(1)$ denotes a sequence of stochastically bounded random variables.

Remark 12. In the next two statements, the EW-TLS estimate $\hat{X}_{\text{EW-TLS}}$ is defined by problem (6) with the additional constraint $\operatorname{rank}(X(\mathscr{J},:)) = l$. We mention that the true value X_0 satisfies this constraint due to condition (iv).

Theorem 13 (*Uniqueness of the estimator*).

(1) *Assume that* (i) *to* (vii) *hold, then*

$$\|\hat{X}_{\text{EW-TLS}} - X_0\|_F = \mu_m \cdot \mathrm{O_p}(1)$$

with

$$\mu_m := \frac{m^{1/4}}{\lambda_{\min}^{1/2}(A_0^\top A_0)} + \frac{\lambda_{\max}^{1/2}(A_0^\top A_0)}{\lambda_{\min}(A_0^\top A_0)}. \tag{21}$$

(2) *Assume that* (i)–(iv), (vi), (viii), *and* (ix) *hold. Then*

(a) *the problem* (12) *has a unique solution* (*and therefore* $\hat{X}_{\text{EW-TLS}}$ *exists and is unique*) *with probability tending to one*;

(b) *there exists a neighborhood* $U_\rho(X_0)$ *such that the equation* $f_0'(X)=0$, $X \in U_\rho(X_0)$ *has a unique solution*, *with probability tending to one*, *and this solution coincides with the estimator* $\hat{X}_{\text{EW-TLS}}$.

Proof. See Appendix B. □

Remark 14. As the condition number of A_0 grows, μ_m also grows. This shows that for ill conditioned problems a larger sample size is needed for accurate estimation.

Theorem 15 (*Local convergence of the computational algorithm*). *Assume that* (i)–(iv), (vi), (viii), *and* (ix) *hold. Then for each confidence probability* $1-\gamma$, *there exists a neighborhood* $U_{\rho_\gamma}(X_0)$ ($\rho_\gamma \leqslant \rho$, ρ *comes from Theorem* 13), *a positive number* C_γ *and an integer* m_γ, *such that for all* $m \geqslant m_\gamma$,

$$\Pr\{\|\Delta X^{(k)}\|_F \leqslant C_\gamma(\mu_m + \|\Delta X^{k-1}\|_F)\,\|\Delta X^{(k-1)}\|_F \;\; for\ k=1,2,\ldots\} \geqslant 1-\gamma,$$

where the initial approximation $X^{(0)} = X^{(0)}(m) \in U_{\rho_\gamma}(X_0)$ *with probability greater than* $1-\gamma/2$, *and* $\Delta X^{(k)} := X^{(k)} - \hat{X}$, *and* μ_m (*tending to zero*) *is given in Theorem* 13. *In particular*, *if* $X^{(0)} \to X_0$, *as* $m \to \infty$, *in probability*, *then*

$$\lim_{m\to\infty} \Pr\left\{\lim_{k\to\infty} \|\Delta X^{(k)}\|_F = 0\right\} = 1.$$

Proof. See Appendix C. □

Theorem 15 states that if the initial approximation is sufficiently close to the EW-TLS estimator $X_{\text{EW-TLS}}$, i.e., to the global minimum point of the minimized cost function f_0, then the algorithm almost surely converges to $X_{\text{EW-TLS}}$. For a fixed sample size, the convergence is linear. Note, however that for large sample size μ_m vanishes, so that the convergence is almost quadratic.

6. Simulation examples

This section shows simulation examples with the EW-TLS estimator. In Section 6.1, we illustrate the consistency of the EW-TLS estimator, and in Section 6.2, we compare the results obtained by TLS, GTLS, and EW-TLS for Example 1 of the introduction.

6.1. Asymptotic behavior of the estimates

We set up a simulation example corresponding to the measurement error model (4) with $n=2$, $l=1$, and m ranging from 75 to 750. The random matrix $\tilde{D} := [\tilde{A}\ \tilde{b}]$ has normal, independent elements with variances $\operatorname{var}(\tilde{D}_{ij}) = \sigma_{ij}^2$. The true data matrix $[A_0\ b_0]$ is random with independent, uniformly distributed elements in the interval [0, 1].

For a fixed $m \in [75, 750]$, $N=500$ noise realizations are generated and the corresponding EW-TLS estimates $\hat{x}(m, N)$ are computed. The relative errors of estimation $e(m, N) := ||\hat{x}(m, N) - x_0||/||x_0||$, are averaged. Let $e(m) := \sum_{i=1}^{N} e(m, i)/N$. The function $e(m)$ is plotted (see Fig. 1) for four noise scenarios defined below.

The matrix of the element-wise specified error standard deviations $\Sigma := [\sigma_{ij}]$ characterizes the experiment. We select Σ in four different ways corresponding to four noise scenarios. Let $i_1 : i_2$ be the set $\{i_1, i_1+1, \ldots, i_2-1, i_2\}$, $\mathbf{1}$ be a vector of ones $[1\ \cdots\ 1]^\top$, and $U(\underline{u}, \overline{u})$ be a matrix of independent and uniformly distributed elements in the interval $[\underline{u}, \overline{u}]$. (The dimensions of $\mathbf{1}$ and $U(\underline{u}, \overline{u})$ are understood from the context.) The four noise scenarios are:

(1) EW-TLS setup—$\Sigma(:, 1:2) = U(0.01, 0.26)$, and $\Sigma(:, 3) = U(0.01, 0.035)$.
(2) WLS setup—$\Sigma(:, 1:2) = 0$ and $\Sigma(:, 3) = U(0.01, 0.51)$.
(3) TLS setup—$\sigma_{ij} = 0.1$ for all $i = 1, \ldots, m$ and $j = 1, 2, 3$.
(4) GTLS setup—$\Sigma = \mathbf{1}u^\top$, where $u \in \mathbb{R}^{m\times 3}$ is $u = U(0.02, 0.52)$.

The computation of the EW-TLS estimator is performed with the proposed algorithm. As initial approximation, in all cases, we use the GTLS estimate.

Convergence of the relative error of estimation to zero as the sample size is increased, indicates consistency of the estimator. The simulation results confirm that the EW-TLS estimator is consistent in the simulated noise setups. In the special cases of the WLS, TLS, and GTLS noise setup, the EW-TLS estimator coincides with the corresponding estimator, which is known to be a consistent for that noise setup. Thus the EW-TLS method is indeed a generalization of the previously known methods.

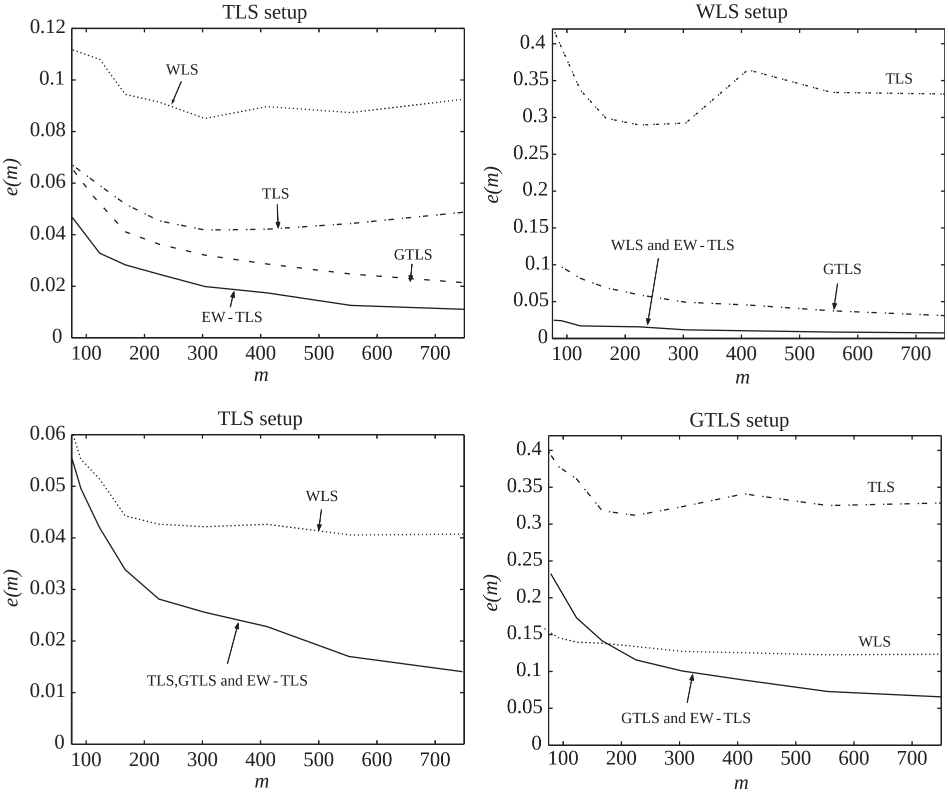

Fig. 1. Relative error of estimation in four noise scenarios, averaged for 500 repetitions.

6.2. Relative error total least-squares

Consider Example 1 from the introduction. In this section, we show simulation results that illustrate the applicability of the problem and compare the TLS, GTLS, and EW-TLS approximations.

The data matrix D is constructed as follows: $m=10$, $n=2$, $l=1$, $A=U(0,1)$, $x=U(0,1)$ (the notation $U(0,1)$ is defined in Section 6.1), $b(2:m)=A(2:m,:)x$, and $b_1=10$. Note that the elements of A are in the interval $[0,1]$ and the elements of $b(2:m)$ are in the interval $[0,2]$. Therefore, the elements of the data matrix D, except for $D_{1,3}$, are small compared to $D_{1,3}=10$. In this case, the TLS method tends to approximate well the large element $D_{1,3}$ and ignore the others. This undesirable effect, is avoided by proper scaling, e.g., proportional with the reciprocal of the size of the elements, which results in the relative error criterion (1).

The weight matrix V has to be chosen in order to apply the GTLS method. We take V diagonal with the ith diagonal element equal to the average of the elements in the ith column of the data matrix D. This results in the most reasonable approximation of the relative error criterion (1) that we aim to minimize.

For a particular simulation example, the matrices of the element-wise relative errors

$$\Delta D_{\mathrm{rel}} := \left[\frac{|d_{ij} - \hat{d}_{ij}|}{d_{ij}} \right]$$

for the TLS, GTLS, and EW-TLS solutions are

$$\Delta D_{\mathrm{rel,tls}} = \begin{bmatrix} 1.5618 & 0.0261 & 0.0001 \\ 0.5668 & 2.8818 & 0.1037 \\ 0.3876 & 0.5493 & 0.0350 \\ 2.5035 & 0.2807 & 0.0246 \\ 8.7380 & 0.3281 & 0.0294 \\ 0.1139 & 0.0768 & 0.0058 \\ 0.5722 & 3.1580 & 0.1081 \\ 12.6720 & 0.3352 & 0.0301 \\ 11.8342 & 0.3340 & 0.0300 \\ 0.2219 & 0.0910 & 0.0073 \end{bmatrix},$$

$$\Delta D_{\mathrm{rel,gtls}} = \begin{bmatrix} 1.4357 & 0.1017 & 0.0038 \\ 0.0921 & 1.9847 & 0.7345 \\ 0.0446 & 0.2678 & 0.1753 \\ 0.7221 & 0.3431 & 0.3091 \\ 2.3755 & 0.3780 & 0.3481 \\ 0.0280 & 0.0800 & 0.0620 \\ 0.0935 & 2.1879 & 0.7701 \\ 3.4188 & 0.3832 & 0.3540 \\ 3.1966 & 0.3824 & 0.3531 \\ 0.1170 & 0.2035 & 0.1688 \end{bmatrix},$$

and

$$\Delta D_{\mathrm{rel,ew\text{-}tls}} = \begin{bmatrix} 0.0002 & 0.0386 & 0.9580 \\ 0.0013 & 0.0009 & 0.0022 \\ 0.0011 & 0.0026 & 0.0037 \\ 0.0002 & 0.0070 & 0.0071 \\ 0.0001 & 0.0073 & 0.0073 \\ 0.0009 & 0.0048 & 0.0057 \\ 0.0016 & 0.0010 & 0.0026 \\ 0.0001 & 0.0074 & 0.0073 \\ 0.0001 & 0.0074 & 0.0073 \\ 0.0007 & 0.0058 & 0.0064 \end{bmatrix}.$$

Note that $\Delta D_{\mathrm{rel,tls},13} = 0.0001$ and $\Delta D_{\mathrm{rel,gtls},13} = 0.0038$ are small but the other elements in these matrices are much larger. This is a numerical demonstration of the above mentioned undesirable effect in using the TLS and GTLS methods for approximation of data with very small and very large elements. (Of course, the example is deliberately chosen to show the effect. For nearly equal size of the elements in D, it is not pronounced.) The corresponding "total" relative errors $||\Delta D_{\mathrm{rel}}||_F^2$, i.e., the values of the cost function of (1), are $||\Delta D_{\text{rel,tls}}||_F^2 = 405$, $||\Delta D_{\text{rel,gtls}}||_F^2 = 41$, and $||\Delta D_{\text{rel,ew-tls}}||_F^2 = 0.92$. The example illustrates the advantage of introducing element-wise scaling in the approximation criterion, in order to achieve adequate approximation.

7. Conclusion

We have formulated a new total least-squares problem that is appropriate for solving overdetermined system of equations with row-wise independent and differently sized errors. Moreover, correlation of the errors in the rows of the extended data matrix and noise-free elements are allowed. The problem is defined as a constrained optimization problem with the parameter estimate and the noise correction as decision variables. We derived an equivalent unconstrained problem in the parameter estimates only. An iterative algorithm is proposed for the latter problem that solves the first-order optimality condition by successive approximations with a linear equation. The algorithm is proven to be locally convergent with linear convergence rate.

Acknowledgements

Dr. Sabine Van Huffel is a full professor and Ivan Markovsky is a research assistant at the Katholieke Universiteit Leuven, Belgium. Dr. Alexander Kukush is supported by a postdoctoral research fellowship of the Belgian office for Scientific, Technical and Cultural Affairs. Research supported by: Research Council KUL: GOA-Mefisto 666, IDO /99/003 and /02/009; Flemish Government: FWO: Ph.D./postdoc Grants, Projects, G.0078.01, G.0407.02, G.0269.02, G.0270.02, research communities (ICCoS, ANMMM); IWT: Ph.D. Grants; Belgian Federal Science Policy Office IUAP P5/22; EU: PDT-COIL, BIOPATTERN, ETUMOUR.

Appendix A. Derivation of $f_0'(X)$

Denote by $\mathscr{D}$ the differential operator. It acts on a differentiable function $f_0 : U \to \mathbb{R}$, where U is an open set in $\mathbb{R}^{n\times l}$ and gives as a result another function, the differential of f_0, $\mathscr{D}(f_0) : U \times \mathbb{R}^{n\times l} \to \mathbb{R}$. $\mathscr{D}(f_0)$ is linear in its second argument, i.e.,

$$\mathscr{D}(f) := \mathrm{d}\, f_0(X, H) = \operatorname{trace}(f_0'(X) H^\top), \tag{A.1}$$

where $f_0' : U \to \mathbb{R}^{n\times l}$ is the derivative of f_0, and has the property

$$f_0(X + H) = f_0(X) + \mathrm{d}\, f_0(X, H) + \mathrm{o}(||H||_F) \tag{A.2}$$

for all $X \in U$ and for all $H \in \mathbb{R}^{n\times l}$. (The notation $\mathrm{o}(||H||_F)$ has the usual meaning: $g(H) = \mathrm{o}(||H||_F)$ if and only if $\lim_{||H||_F \to 0} g(H)/||H||_F = 0$.)

Let

$$Q_i(X) := X_{\text{ext}}^\top V_{\tilde{d}_i} X_{\text{ext}},$$

so that

$$f_0(X) = \sum_{i=1}^{m} r_i^\top(X) Q_i^{-1}(X) r_i(X).$$

We find the derivative $f_0'(X)$ by first deriving the differential $\mathscr{D}(f_0)$ and then representing it in the form (A.1) from which $f_0'(X)$ is extracted. The differential of f_0 is

$$\begin{aligned} &\mathrm{d}\, f_0(X, H) \\ &\quad = \sum_{i=1}^{m} \left(a_i^\top H Q_i^{-1}(X) r_i(X) + r_i^\top(X) Q_i^{-1}(X) H^\top a_i + r_i^\top(X) \mathscr{D}(Q_i^{-1}(X)) r_i(X) \right) \\ &\quad = \sum_{i=1}^{m} \left(2\mathrm{trace}(a_i r_i^\top(X) Q_i^{-1}(X) H^\top) + \mathrm{trace}\left(\mathscr{D}(Q_i^{-1}(X)) r_i(X) r_i^\top(X) \right) \right). \end{aligned}$$

Using the rule for differentiation of an inverse matrix-valued function, we have

$$\mathscr{D}(Q_i^{-1}(X)) \equiv -Q_i^{-1}(X) \mathscr{D}(Q_i(X)) Q_i^{-1}(X).$$

Using the defining property (A.2), we have

$$\begin{aligned} \mathscr{D}(Q_i(X)) \equiv \mathscr{D}\left([X \ \ -I] \ V_{\tilde{d}_i} \begin{bmatrix} X \\ -I \end{bmatrix} \right) &= \mathrm{trace}\left([H^\top \ \ 0] \, V_{\tilde{d}_i} \begin{bmatrix} X \\ -I \end{bmatrix} + [X \ \ -I] V_{\tilde{d}_i} \begin{bmatrix} H \\ 0 \end{bmatrix} \right) \\ &= 2\mathrm{trace}\left([H^\top \ \ 0] \, V_{\tilde{d}_i} \begin{bmatrix} X \\ -I \end{bmatrix} \right). \end{aligned}$$

The covariance matrix $V_{\tilde{d}_i}$ is partitioned as

$$V_{\tilde{d}_i} = \begin{bmatrix} \mathrm{cov}(\tilde{a}_i) & \mathrm{cov}(\tilde{a}_i, \tilde{b}_i) \\ \mathrm{cov}(\tilde{b}_i, \tilde{a}_i) & \mathrm{cov}(\tilde{b}_i) \end{bmatrix} =: \begin{bmatrix} V_{\tilde{a}_i} & V_{\tilde{a}_i \tilde{b}_i} \\ V_{\tilde{b}_i \tilde{a}_i} & V_{\tilde{b}_i} \end{bmatrix}.$$

Then

$$\mathscr{D}(Q_i(X)) \equiv 2\mathrm{trace}(H^\top (V_{\tilde{a}_i} X - V_{\tilde{a}_i \tilde{b}_i})).$$

Substituting backwards, we have

$$\begin{aligned} \mathrm{d}\, f_0(X, H) = &\sum_{i=1}^{m} \Big(2\mathrm{trace}(a_i r_i^\top(X) Q_i^{-1}(X) H^\top) \\ &-2\mathrm{trace}(Q_i^{-1}(X) H^\top (V_{\tilde{a}_i} X - V_{\tilde{a}_i \tilde{b}_i}) Q_i^{-1}(X) r_i(X) r_i^\top(X)) \Big) \\ = \mathrm{trace} &\left(\left(2 \sum_{i=1}^{m} \Big(a_i r_i^\top(X) Q_i^{-1}(X) \right.\right. \\ &\left.\left. -(V_{\tilde{a}_i} X - V_{\tilde{a}_i \tilde{b}_i}) Q_i^{-1}(X) r_i(X) r_i^\top(X) Q_i^{-1}(X) \Big) \right) H^\top \right). \end{aligned}$$

Thus

$$f_0'(X) = 2\sum_{i=1}^{m}\left(a_i r_i^\top(X)Q_i^{-1}(X) - (V_{\tilde{a}_i}X - V_{\tilde{a}_i\tilde{b}_i})Q_i^{-1}(X)r_i(X)r_i^\top(X)Q_i^{-1}(X)\right).$$

Appendix B. Proof of Theorem 13

First, we prove part 1 *of the theorem*: From Kukush and Van Huffel (2004), see the proof of Theorem 2, we have for $\Delta\hat{X} := \hat{X} - X_0$, and for each $\zeta > 0$, that

$$\Pr\{||\Delta\hat{X}||_F > \zeta\} \leqslant \text{const}\left(1 + \frac{1}{\zeta^{2\delta}}\right)\mu_m^{2\delta},$$

where μ_m is given in (21). Therefore, for each $c > 0$, we have

$$\Pr\{||\Delta\hat{X}||_F > c\mu_m\} \leqslant \text{const}\left(\mu_m^{2\delta} + \frac{1}{c^{2\delta}}\right).$$

By assumptions (vi) and (vii) we have $\mu_m \to 0$, as $m \to \infty$, therefore

$$\limsup_{m\to\infty} \Pr\left\{\frac{||\Delta\hat{X}||_F}{\mu_m} > c\right\} \leqslant \text{const}\,\frac{1}{c^{2\delta}}$$

and

$$\lim_{c\to+\infty}\limsup_{m\to\infty} \Pr\left\{\frac{||\Delta\hat{X}||_F}{\mu_m} > c\right\} = 0.$$

This proves that $||\Delta\hat{X}||_F/\mu_m = \mathrm{O_p}(1)$.

Next, we prove part 2: Hereafter "w.p.t.o." stands for "*w*ith *p*robability *t*ending to *o*ne". The derivative $f_0'(X)$ is a symmetric bilinear form on $\mathbb{R}^{n\times d}$. We will show that in a certain neighborhood $U_\rho(X_0)$, w.p.t.o., $f_0'(X)(H,H) \geqslant \text{const}||H||_F^2$, $H \in \mathbb{R}^{n\times d}$, with certain positive constant "const". This implies the uniqueness of the solution of (17), w.p.t.o. From the weak consistency result, see Kukush and Van Huffel (2004), we have that a minimum point of $f_0(X)$ belongs to $U_\rho(X_0)$, w.p.t.o., and it is a root of (17). Therefore the solution of equation (17) is not only unique, but it exists and coincides with $\hat{X}$, w.p.t.o. Since $\Pr\{\hat{X} \notin U_\rho(X_0)\} \to 0$, as $m \to \infty$, w.p.t.o. $\hat{X} \in U_\rho(X_0)$, and $\hat{X}$ is unique w.p.t.o.

Therefore to prove the statements it is enough to construct $U_\rho(X_0)$ such that w.p.t.o., for a certain positive constant "const",

$$F'(X)(H,H) \geqslant \text{const}||H||_F^2, \quad X \in U_\rho(X_0), \quad H \in \mathbb{R}^{n\times d}. \tag{B.1}$$

After straightforward but tedious calculations, see (Kukush et al., 2002, Section 7.1) for the complete derivation, we have,

$$\begin{aligned}\operatorname{trace}(H^\top f_0' H) = \sum_{i=1}^{m} \operatorname{trace}\Big(& a_i a_i^\top H Q_i^{-1} H^\top - V_{\tilde a_i} H Q_i^{-1} S_i Q_i^{-1} H^\top \\ & - 2 a_i (a_i^\top X - b_i^\top) Q_i^{-1} (H^\top W_i + W_i^\top H) Q_i^{-1} H^\top \\ & + 2 Q_i^{-1} W_i^\top H Q_i^{-1} S_i Q_i^{-1} W_i^\top H + 2 Q_i^{-1} W_i^\top H Q_i^{-1} S_i Q_i^{-1} H^\top W_i \Big),\end{aligned}$$

where

$$W_i := V_{\tilde a_i} X - V_{\tilde a_i \tilde b_i} \quad \text{and} \quad S := \left(d_i^\top \begin{bmatrix} X \\ -I \end{bmatrix} \right)^\top \left(d_i^\top \begin{bmatrix} X \\ -I \end{bmatrix} \right).$$

Now, we compute the expectation of $\operatorname{trace}(H^\top f_0' H)$. For some positive constants “const_i” and for $\|\Delta X\|_F \leqslant \rho_1$, $\Delta X := X - X_0$, see (Kukush et al., 2002, Section 7.2),

$$\begin{aligned}\mathbf{E} \operatorname{trace}(H^\top f_0' H) \geqslant (& \text{const}_1 \lambda_{\min} - \text{const}_2 \lambda_{\max} \|\Delta X\|_F \\ & - \text{const}_3 \lambda_{\max} \|\Delta X\|_F^2) \|H\|_F^2. \end{aligned} \tag{B.2}$$

Hereafter, for brevity we denote $\lambda_{\min}(A_0^\top A_0)$ by $\lambda_{\min}$ and $\lambda_{\max}(A_0^\top A_0)$ by $\lambda_{\max}$. Due to assumption (ix), (B.2) yields for $\|\Delta X\|_F \leqslant \rho_2$ that

$$\mathbf{E} \operatorname{trace}(H^\top f_0' H) \geqslant \text{const}_4 \lambda_{\min} \|H\|_F^2 \tag{B.3}$$

and the bound (B.1) is obtained.

Now, we analyze $H^\top f_0' H - \mathbf{E}(H^\top f_0' H)$. We have for instance

$$a_i a_i^\top - \mathbf{E}(a_i a_i^\top) = \tilde a_i a_{0,i}^\top + a_{0,i} \tilde a_i^\top + (\tilde a_i \tilde a_i^\top - V_{\tilde a_i}) =: P_1$$

and the corresponding summand in $H^\top f_0' H - \mathbf{E}(H^\top f_0' H)$ is

$$S_1(X) = \sum_{i=1}^{m} P_{1,i} H Q_i^{-1} H^\top.$$

To bound this sum we use a matrix version of Rosenthal inequality, see, e.g., Lemma 2 in Kukush and Van Huffel (2004). As δ in (viii) is greater than or equal to 2, we have

$$\mathbf{E}\|S_1(X_0)\|_F^\delta \leqslant \text{const} \|H\|_F^{2\delta} (\lambda_{\max}^{\delta/2} + m^{\delta/2})$$

and for $X_1, X_2 \in \Theta := \{X : \|X - X_0\|_F \leqslant \rho_2\}$ we have

$$\mathbf{E}\|S_1(X_1) - S_1(X_2)\|^\delta \leqslant \text{const} \|H\|_F^{2\delta} (\lambda_{\max}^{\delta/2} + m^{\delta/2}) \|X_1 - X_2\|^\delta.$$

By condition (viii), $\delta > nd$, and we apply Lemma 1 from Kukush and Van Huffel (2004) to the compact set Θ. We have for any $c > 0$,

$$\Pr\left\{ \sup_{\|X - X_0\|_F \leqslant \rho_2} \|S_1(X)\| > c \right\} \leqslant \text{const} \frac{\|H\|_F^{2\delta} (\lambda_{\max}^{1/2} + m^{1/2})^\delta}{c^\delta}.$$

This implies that

$$\sup_{||X-X_0||_F \leqslant \rho_2} ||S_1(X)|| = ||H||_F^2 (\lambda_{\max}^{1/2} + m^{1/2}) \mathrm{O_p}(1).$$

The same bounds can be obtained for the other summands of $H^\top f_0' H - \mathbf{E}(H^\top f_0' H)$. Therefore, see (B.3),

$$\begin{aligned}\inf_{||X-X_0||_F \leqslant \rho_2} \operatorname{trace}(H^\top f_0''(X) H) &\geqslant (\mathrm{const}_4 \lambda_{\min} - (\lambda_{\max}^{1/2} + m^{1/2}) \mathrm{O_p}(1)) ||H||_F^2 \\ &= \left(\mathrm{const}_4 - \frac{\lambda_{\max}^{1/2} + m^{1/2}}{\lambda_{\min}} \mathrm{O_p}(1) \right) ||H||_F^2.\end{aligned}$$

By assumptions (vi) and (ix), $(\lambda_{\max}^{1/2} + m^{1/2})/\lambda_{\min} \to 0$, as $m \to \infty$, therefore (B.1) holds in $U_{\rho_1}(X_0)$, w.p.t.o. □

Appendix C. Proof of Theorem 15

We analyze the partial derivatives of F given in (19), *when X varies in a certain* $U_\rho(X_0)$: We have

$$\operatorname{trace}\left(H_2^\top \frac{\partial F}{\partial X} H_1 \right) = \sum_{i=1}^{m} \operatorname{trace}(a_i a_i^\top H_1 Q_{y,i}^{-1} H_2^\top - V_{\tilde{a}_i} H_1 Q_{y,i}^{-1} S_{y,i} Q_{y,i}^{-1} H_2^\top) \quad \text{(C.1)}$$

with $S_{y,i} := (Y^\top a_i - b_i)(a_i^\top Y - b_i^\top)$. But (C.1) is a symmetric bilinear form on $\mathbb{R}^{n \times d}$. Therefore we can identify $\partial F / \partial X$ with a self-adjoint operator on $\mathbb{R}^{n \times d}$, it is uniquely characterized by the corresponding quadratic form. Then

$$\operatorname{trace}\left(H^\top \frac{\partial F}{\partial X} H \right) = \sum_{i=1}^{m} \operatorname{trace}(a_i a_i^\top H Q_{y,i}^{-1} H^\top - V_{\tilde{a}_i} H Q_{y,i}^{-1} S_{y,i} Q_{y,i}^{-1} H^\top)$$

and with $\Delta Y := Y - X_0$,

$$\begin{aligned}\mathbf{E} \operatorname{trace}\left(H^\top \frac{\partial F}{\partial X} H \right) = \sum_{i=1}^{m} &\operatorname{trace}(a_{0,i} a_{0,i}^\top H Q_{y,i}^{-1} H^\top \\ &- V_{\tilde{a}_i} H Q_{y,i}^{-1} \Delta Y^\top a_{0,i} a_{0,i}^\top \Delta Y Q_{y,i}^{-1} H^\top).\end{aligned}$$

As in Appendix B, we have for $||\Delta Y||_F \leqslant \rho$ that

$$\mathbf{E} \operatorname{trace}\left(H^\top \frac{\partial \Phi}{\partial X} H \right) \geqslant \mathrm{const}_5 \lambda_{\min} ||H||_F^2 - \mathrm{const}_6 ||\Delta Y||_F^2 \lambda_{\max} ||H||_F^2,$$

and under assumption (ix), for $||\Delta Y||_F \leqslant \rho_3$, we have

$$\mathbf{E} \operatorname{trace}\left(H^\top \frac{\partial \Phi}{\partial X} H \right) \geqslant \mathrm{const}_7 \lambda_{\min} ||H||_F^2.$$

Similarly to Appendix B, we have

$$\sup_{||Y-X_0||\leqslant\rho_3} \left\| H^\top \frac{\partial F}{\partial X} H - H^\top \mathbf{E}\, \frac{\partial F}{\partial X} H \right\|_F = ||H||_F^2 (\lambda_{\max}^{1/2} + m^{1/2}) \mathrm{O_p}(1).$$

Then

$$\inf_{||Y-X_0||\leqslant\rho_3} \operatorname{trace}\left(H^\top \frac{\partial F}{\partial X} H \right) \geqslant \lambda_{\min} \left(\mathrm{const}_7 - \frac{\lambda_{\max}^{1/2} + m^{1/2}}{\lambda_{\min}} \mathrm{O_p}(1) \right) ||H||_F^2.$$

Therefore the linear equation (19) has a unique solution $X = \Psi(Y)$, where $||Y - X_0|| \leqslant \rho_3$, w.p.t.o.

The function F has the following structure:

$$F(X, Y) = M(Y)X + Q(Y). \tag{C.2}$$

Here, $M(Y) = \partial F/\partial X$ is a linear operator in $\mathbb{R}^{n\times d}$ at a fixed point $Y \in \mathbb{R}^{n\times d}$, and $Q(Y) \in \mathbb{R}^{n\times d}$. Note that we have redefined Q introduced before.

The solution of (19) is

$$X = -M(Y)^{-1} Q(Y) = \Psi(Y). \tag{C.3}$$

Here M^{-1} is an inverse operator in $\mathbb{R}^{n\times d}$. We study the behavior of M, Q, its derivatives and finally of $\Psi'(Y)$ in a neighborhood of X_0.

Due to the analysis above for $||\Delta Y||_F \leqslant \rho_3$, we have

$$M(Y)X = \sum_{i=1}^{m} (a_{0,i} a_{0,i}^\top X Q_{u,i}^{-1} - \Sigma_{a_i} X Q_{u,i}^{-1} \Delta Y^\top a_{0,i} a_{0,i}^\top \Delta Y Q_{u,i}^{-1}) + R_1(Y)X \tag{C.4}$$

with $||R_1(Y)|| = (\lambda_{\max}^{1/2} + m^{1/2}) \mathrm{O_p}(1)$.

Let $W_{y,i} := V_{\tilde{a}_i} Y - V_{\tilde{a}_i \tilde{b}_i}$. We have, see (Kukush et al., 2002, Section 8.2), that

$$\begin{aligned} \mathbf{E}\, \frac{\partial (M(Y)X)}{\partial Y} H = & -\sum_{i=1}^{m} a_{0,i} a_{0,i}^\top X Q_{u,i}^{-1} (H^\top W_{u,i} + W_{u,i}^\top H) Q_{u,i}^{-1} \\ & + R_2(Y)(X, H) + R_3(Y)(X, H) \end{aligned}$$

with

$$||R_3(Y)(X, H)|| = (\lambda_{\max}^{1/2} + m^{1/2}) ||X|| ||H|| \mathrm{O_p}(1)$$

and

$$||R_2(Y)(X, H)|| \leqslant \lambda_{\max} ||X|| ||H|| ||\Delta Y|| \mathrm{const}.$$

Also, see (Kukush et al., 2002, Section 8.3),

$$Q(Y) = -\sum_{i=1}^{m} a_{0,i} a_{0,i}^\top X_0 Q_{u,i}^{-1} + R_4(Y)$$

and in $U_{\rho_3}(X_0)$

$$||R_4(Y)|| \leqslant (\lambda_{\max}^{1/2} + m^{1/2})\mathrm{O_p}(1) + \lambda_{\max}||\Delta Y||_F^2 \text{const.}$$

Finally,

$$\mathbf{E}\left(\frac{\partial Q(Y)}{\partial Y}\right) H = \sum_{i=1}^{m} a_{0,i} a_{0,i}^{\top} X_0 Q_u^{-1} (H^{\top} W_u + W_u^{\top} H) Q_u^{-1} + R_5(Y) H$$

with

$$||R_5(Y)|| \leqslant \lambda_{\max} ||\Delta Y|| \text{const.}$$

Then

$$\frac{\partial Q(Y)}{\partial Y} H = \sum_{i=1}^{m} a_{0,i} a_{0,i}^{\top} X_0 Q_{u,i}^{-1} (H^{\top} W_{u,i} + W_{u,i}^{\top} H) Q_{u,i}^{-1} + R_5(Y) H + R_6(Y) H \tag{C.5}$$

with

$$||R_6(Y)|| = (\lambda_{\max}^{1/2} + m^{1/2})\mathrm{O_p}(1).$$

Next, we analyze the derivative of Ψ, given in (C.3). We have for $H \in \mathbb{R}^{n\times d}$

$$\frac{\partial \Psi}{\partial Y} H = -M^{-1} \frac{\partial Q}{\partial Y} H + M^{-1} \left.\frac{\partial (MZ)}{\partial Y}\right|_{Z=M^{-1}Q} H.$$

Now, we use (C.4) and (C.5) to obtain

$$\frac{\partial \Psi}{\partial Y} H = M^{-1} \left(-\frac{\partial Q}{\partial Y} H + \left.\frac{\partial (MZ)}{\partial Y}\right|_{Z=M^{-1}Q} H \right).$$

Denote

$$T_{XY} := \sum_{i=1}^{m} a_{0,i} a_{0,i}^{\top} X Q_{u,i}^{-1}, \quad S_{XY} H := \sum_{i=1}^{m} a_{0,i} a_{0,i}^{\top} X Q_u^{-1} (H^{\top} W_y + W_y^{\top} H) Q_y^{-1}.$$

Then

$$Q(Y) = -T_{X_0 Y} + R_4, \quad \frac{\partial Q}{\partial Y} = S_{X_0 Y} + R_5', \tag{C.6}$$

$$M(Y)X = T_{XY} + R_1', \quad \frac{\partial (M(Y)Z)}{\partial Y} = -S_{ZY} + R_2'. \tag{C.7}$$

Here for $Y \in U_{\rho_3}(X_0)$,

$$||R_4(Y)|| = (\lambda_{\max}^{1/2} + m^{1/2})\mathrm{O_p}(1) + \lambda_{\max}||\Delta Y||_F^2 \mathrm{O}(1),$$

$$||R'_5(Y)|| = (\lambda_{\max}^{1/2} + m^{1/2})\mathrm{O_p}(1) + \lambda_{\max}||\Delta Y||_F \mathrm{O}(1),$$

$$||R'_1(Y, X)|| \leqslant ||X|| \left((\lambda_{\max}^{1/2} + m^{1/2})\mathrm{O_p}(1) + \lambda_{\max}||\Delta Y||_F^2 \mathrm{O}(1)\right),$$

$$||R'_2(Y, Z)|| \leqslant ||Z|| \left((\lambda_{\max}^{1/2} + m^{1/2})\mathrm{O_p}(1) + \lambda_{\max}||\Delta Y||_F \mathrm{O}(1)\right).$$

We explain why $||\partial\Psi/\partial Y H||$ is small enough: If we neglect the residuals, then

$$\frac{\partial\Psi}{\partial Y} H \approx M^{-1}(-S_{X_0 Y} H - S_{ZY} H),$$

here Z is found from $M(Y)Z \approx Q(Y)$, or $T_{ZY} \approx -T_{X_0}$, $Z \approx -X_0$. Then

$$\frac{\partial\Psi}{\partial Y} H \approx M^{-1}(-S_{X_0 Y} H - S_{(-X_0, Y)} H) = 0.$$

Now, we give a bound for $||\partial\Psi/\partial Y||$: The operator $M(Y) = \partial F/\partial X$ is a positive self-adjoint operator in $\mathbb{R}^{d\times d}$, w.p.t.o., and due to (19)

$$\inf_{||Y - X_0|| \leqslant \rho_3} \lambda_{\min}(M(Y)) \geqslant \lambda_{\min}\left(\mathrm{const}_7 - \frac{\lambda_{\max}^{1/2} + m^{1/2}}{\lambda_{\min}} \mathrm{O_p}(1)\right).$$

Therefore for $Y \in B(X_0, \rho_3)$ w.p.t.o.,

$$||M^{-1}(Y)|| = \frac{1}{\lambda_{\min}(M(Y))}.$$

Then we have

$$\left|\left|\frac{\partial\Psi}{\partial Y}\right|\right|_F \leqslant \frac{1}{\lambda_{\min}(M(Y))} \left|\left| -\frac{\partial Q}{\partial Y} + \left.\frac{\partial (MZ)}{\partial Y}\right|_{Z = M^{-1}Q} \right|\right| \tag{C.8}$$

and due to the bounds for the residuals, we have

$$\left|\left|\frac{\partial\Psi}{\partial Y}\right|\right| \leqslant \frac{\lambda_{\max}||\Delta Y||_F \mathrm{O_p}(1) + (\lambda_{\max}^{1/2} + m^{1/2})\mathrm{O_p}(1)}{\lambda_{\min}}. \tag{C.9}$$

The last inequality can be explained in more detail. Consider the norm on the right-hand side of (C.8).

$$\begin{aligned} -\frac{\partial Q}{\partial Y} + \left.\frac{\partial (MZ)}{\partial Y}\right|_{Z = M^{-1}Q} &= -S_{X_0 Y} + R'_5 - S_{ZY} + R'_2 \\ &= S_{-Z - X_0, Y} + R'_5 + R'_2. \end{aligned} \tag{C.10}$$

Now,

$$\begin{aligned} &MZ = Q, \quad T_{ZY} + R'_1(Z) = -T_{X_0 Y} + R_4(Y), \\ &T_{-Z - X_0, Y} + R'_1(-Z - X_0) = -R_4(Y) - R'_1(X_0), \end{aligned}$$

Here $R_1'(X)$ is linear in X; it depends also on Y, but we suppress that dependence.

$$M(Y)(-Z-X_0) = -R_4(Y) - R_1'(X_0), \quad -Z-X_0 = -M^{-1}(R_4(Y)+R_1'(X_0))$$

and then

$$||-Z-X_0||_F \leqslant \frac{1}{\lambda_{\min}}((\lambda_{\max}^{1/2}+m^{1/2})\mathrm{O_p}(1)+\lambda_{\max}||\Delta Y||_F^2\mathrm{O}(1)).$$

Therefore

$$||S_{-Z-X_0,Y}|| \leqslant \frac{\lambda_{\max}}{\lambda_{\min}}\left((\lambda_{\max}^{1/2}+m^{1/2})\mathrm{O_p}(1)+\lambda_{\max}||\Delta Y||_F^2\mathrm{O}(1)\right).$$

From the last inequality, (C.10) and (C.8), we obtain (C.9).

Due to assumption (ix), we have

$$\left|\left|\frac{\partial \Psi}{\partial Y}\right|\right| \leqslant (||\Delta Y||_F + \mu_m)\mathrm{O_p}(1). \tag{C.11}$$

Here, we used the equation $(\lambda_{\max}^{1/2}+m^{1/2})/\lambda_{\min} = \mu_m \mathrm{O}(1)$.

We obtain the rate of convergence: Let $1-\gamma$ be a confidence probability. Then there exists a positive real c_γ such that for $\mathrm{O_p}(1)$ in (C.11),

$$\sup_{m\geqslant m_0} \Pr\{\mathrm{O_p}(1) > c_\gamma\} \leqslant \gamma.$$

Here, m_0 is large enough but a fixed number. Then for every $m \geqslant m_0$, with probability greater than or equal to $1-\gamma$,

$$\left|\left|\frac{\partial \Psi}{\partial Y}\right|\right| \leqslant (||\Delta Y||_F + \mu_m)c_\gamma.$$

Now, ρ_γ is chosen from the condition

$$\rho_\gamma \leqslant \rho_3, \quad \rho_\gamma c_\gamma \leqslant \tfrac{1}{2} \quad \text{and} \quad \rho_\gamma \leqslant \frac{1}{2c_\gamma}.$$

Then for all $m \geqslant m_\gamma$, with certain m_γ, we have $\mu_m c_\gamma \leqslant \frac{1}{4}$. Therefore each $Y \in \bar{U}_{\rho_\gamma}(X_0)$, $m \geqslant m_\gamma$,

$$\left|\left|\frac{\partial \Psi}{\partial Y}\right|\right| \leqslant \frac{3}{4}. \tag{C.12}$$

Here $\bar{U}_{\rho_\gamma}(X_0) := \{X : ||X-X_0||_F \leqslant \rho_\gamma\}$.

We want to ensure that Ψ is a mapping from $\bar{U}_{\rho_\gamma}(X_0)$ into itself. We may and do assume that

$$\hat{X} \in \bar{U}_{\rho_\gamma}(X_0), \quad m \geqslant m_\gamma.$$

Due to $\Psi(\hat{X}) = \hat{X}$ and (C.12)

$$||\Psi(Y) - \hat{X}||_F \leqslant \tfrac{3}{4}||Y-\hat{X}||_F.$$

Thus

$$\begin{aligned}||\Psi(Y) - X_0||_F &\leqslant \tfrac{3}{4}||Y - \hat{X}||_F + ||\hat{X} - X_0||_F\\ &\leqslant \tfrac{3}{4}||Y - X_0||_F + \tfrac{7}{4}||\hat{X} - X_0||_F\\ &\leqslant \tfrac{3}{4}\rho_\gamma + \tfrac{7}{4}||\hat{X} - X_0||_F\end{aligned}$$

and for $m \geqslant \tilde{m}_\gamma$, $\Pr(||\hat{X} - X_0||_F \leqslant \frac{1}{4}\rho_\gamma) \geqslant 1 - \gamma$. With probability at least $1 - 2\gamma$, for $m \geqslant \tilde{m}_\gamma$ we have

$$||\Psi(Y) - X_0||_F \leqslant \frac{\rho_\gamma}{4}.$$

Then with probability at least $1 - 2\gamma$, Ψ is a contraction on $\bar{U}_{\rho_\gamma}(X_0)$. If $X^{(0)} \in U_{\rho_\gamma}$, then $\lim_{k\to\infty} X^{(k)} = \hat{X}$, with probability at least $1 - 2\gamma$. Moreover,

$$\begin{aligned}||X^{(k)} - \hat{X}||_F &\leqslant c_\gamma(\mu_m + ||X^{(k-1)} - X_0||_F)||X^{(k-1)} - \hat{X}||_F\\ &\leqslant c_\gamma(\mu_m + \mu_m \mathrm{O_p}(1) + ||X^{(k-1)} - \hat{X}||_F)||X^{(k-1)} - \hat{X}||_F.\end{aligned}$$

Here, Theorem 13 part 1 is used. Then Theorem 15 follows.

References

Cheng, C., Van Ness, J.W., 1999. Statistical Regression with Measurement Error, Arnold, London.

De Moor, B., 1993. Structured total least squares and L_2 approximation problems. Linear Algebra Appl. 188–189, 163–207.

Fisher, G.W., 1989. Matrix analysis of metamorphic mineral assemblages and reactions. Contrib. Mineral. Petrol. 102, 69–77.

Fuller, W.A., 1987. Measurement Error Models, Wiley, New York.

Gallo, P.P., 1982. Consistency of regression estimates when some variables are subject to error. Comm. Statist. B–Theory Methods 11, 973–983.

Golub, G.H., Van Loan, C.F., 1980. An analysis of the total least squares problem. SIAM J. Numer. Anal. 17, 883–893.

Kukush, A., Van Huffel, S., 2004. Consistency of elementwise-weighted total least squares estimator in a multivariate errors-in-variables model $AX = B$. Metrika 59 (1), 75–97.

Kukush, A., Markovsky, I., Van Huffel, S., 2002. About the convergence of the computational algorithm for the EW-TLS estimator. Technical Report 02–49, Department of EE, K.U. Leuven. Available at: ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/markovsky

Premoli, A., Rastello, M.L., 2002. The parametric quadratic form method for solving TLS problems with elementwise weighting. In: Van Huffel, S., Lemmerling, P. (Eds.), Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications. Kluwer, Dordrecht, pp. 67–76.

Sprent, P., 1966. A generalized least-squares approach to linear functional relationships. J. Roy. Statist. Soc. B 28, 278–297.

Van Huffel, S. (Ed.), 1997. Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling. SIAM, Philadelphia.

Van Huffel, S., Lemmerling, P. (Eds.), 2002. Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Van Huffel, S., Vandewalle, J., 1989. Analysis and properties of the generalized total least squares problem $AX \approx B$ when some or all columns in A are subject to error. SIAM J. Matrix Anal. 10 (3), 294–315.

Van Huffel, S., Vandewalle, J., 1991. The Total Least Squares Problem: Computational Aspects and Analysis, SIAM, Philadelphia.