Modern Stochastics: Theory and Applications 5 (3) (2018) 247–295
https://doi.org/10.15559/18-VMSTA104

Consistency of the total least squares estimator in the linear errors-in-variables regression

Sergiy Shklyar

Taras Shevchenko National University of Kyiv

shklyar@univ.kiev.ua (S.V. Shklyar)

Received: 31 October 2017, Revised: 7 May 2018, Accepted: 8 May 2018,
Published online: 30 May 2018

Abstract This paper deals with a homoskedastic errors-in-variables linear regression model and properties of the total least squares (TLS) estimator. We partly revise the consistency results for the TLS estimator previously obtained by the author [18]. We present complete and comprehensive proofs of consistency theorems. A theoretical foundation for construction of the TLS estimator and its relation to the generalized eigenvalue problem is explained. Particularly, the uniqueness of the estimate is proved. The Frobenius norm in the definition of the estimator can be substituted by the spectral norm, or by any other unitarily invariant norm; then the consistency results are still valid.

Keywords Errors in variables, functional model, linear regression, measurement error model, multivariate regression, total least squares, strong consistency

2010 MSC 62J05, 62H12

1 Introduction

We consider a functional linear error-in-variables model. Let $\{a_i^0,\ \ i \geq 1\}$ be a sequence of unobserved nonrandom n-dimensional vectors. The elements of the vectors are true explanatory variables or (in other terminology) true regressors. We observe m n-dimensional random vectors $a_1, \dots, a_m$ and m d-dimensional random vectors $b_1, \dots, b_m$. They are thought to be true vectors a_i^0 and $X_0^\top a_i^0$, respectively, plus additive errors:

$$\begin{cases} b_i = X_0^\top a_i^0 + \tilde{b}_i, \\ a_i = a_i^0 + \tilde{a}_i, \end{cases} \tag{1}$$

where $\tilde{a}_i$ and $\tilde{b}_i$ are random measurement errors in the regressor and in the response. A nonrandom matrix X_0 is estimated based on observations $a_i, b_i, i = 1, \dots, m$.

This problem is related to finding an approximate solution to incompatible linear equations ("overdetermined" linear equation, because the number of equations exceeds the number of variables)

$$AX \approx B,$$

where $A = [a_1, \dots, a_m]^\top$ is an $m \times n$ matrix and $B = [b_1, \dots, b_m]^\top$ is an $m \times d$ matrix. Here X is an unknown $n \times d$ matrix.

In the linear error-in-variables regression model (1), the *Total Least Squares* (TLS) estimator in widely used. It is a multivariate equivalent to the orthogonal regression estimator. We are looking for conditions that provide consistency or strong consistency of the estimator. It is assumed (for granted) that the measurement errors $\tilde{c}_i = \binom{\tilde{a}_i}{\tilde{b}_i}$, $i = 1, 2, \dots$, are independent and have the same covariance matrix $\varSigma$. It may be singular. In particular, some of regressors may be observed without errors. (If the matrix $\varSigma$ is nonsingular, the proofs can be simplified.) An intercept can be introduced into (1) by augmenting the model and inserting a constant error-free regressor.

Sufficient conditions for consistency of the estimator are presented in Gleser [5], Gallo [4], Kukush and Van Huffel [10]. In [18], the consistency results are obtained under less restrictive conditions than in [10]. In particular, there is no requirement that

$$\frac{\lambda^2_{\min}(A_0^\top A_0)}{\lambda_{\max}(A_0^\top A_0)} \to \infty \quad \text{as} \quad m \to \infty,$$

where $A_0 = [a_1^0, \dots, a_m^0]^\top$ is the matrix A without measurement errors. Hereafter, $\lambda_{\min}$ and $\lambda_{\max}$ denotes the minimum and maximum eigenvalues of a matrix if all the eigenvalues are real numbers. The matrix $A_0^\top A_0$ is symmetric (and positive semidefinite). Hence, its eigenvalues are real (and nonnegative).

The model where some variables are explanatory and the other are response is called *explicit*. The alternative is the *implicit* model, where all the variables are treated equally. In the *implicit* model, the n-dimensional linear subspace in $\mathbb{R}^{n+d}$ is fitted to an observed set of points. Some n-dimensional subspaces can be represented in a form $\{(a, b) \in \mathbb{R}^{n+d} : b = X^\top a\}$ for some $n \times d$ matrix X; such subspaces are called *generic*. The other subspaces are called *non-generic*. The true points lie on a generic subspace $\{(a, b) : b = X_0^\top a\}$. A consistently estimated subspace must be generic with high probability. We state our results for the explicit model, but use the ideas of the implicit model in the definition of the estimator, as well as in proofs.

We allow errors in different variables to correlate. Our problem is a minor generalization of the mixed LS-TLS problem, which is studied in [20, Section 3.5]. In the latter problem, some explanatory variables are observed without errors; the other explanatory variables and all the response variables are observed with errors. The errors have the same variance and are uncorrelated. The basic LS model (where the explanatory variables are error-free, and the response variables are error-ridden) and the basic TLS model (where all the variables are observed with error, and the errors are uncorrelated) are marginal cases of the mixed LS-TLS problem. By a linear transformation of variables our model can be transformed into either a mixed LS-TLS or basic LS or basic TLS problem. (We do not handle the case where there are more error-free variables than explanatory variables.) Such a transformation does not always preserve the sets of generic and non-generic subspaces. The mixed LS-TLS problem can be transformed into the basic TLS problem as it is shown in [6].

The Weighted TLS and Structured TLS estimators are generalizations of the TLS estimator for the cases where the error covariance matrices do not coincide for different observations or where the errors for different observations are dependent; more precisely, the independence condition is replaced with the condition on the "structure of the errors". The consistency of these estimators is proved in Kukush and Van Huffel [10] and Kukush et al. [9]. Relaxing conditions for consistency of the Weighted TLS and Structured TLS estimators is an interesting topic for a future research. For generalizations of the TLS problem, see the monograph [13] and the review [12].

In the present paper, for a multivariate regression model with multiple response variables we consider two versions of the TLS estimator. In these estimators, different norms of the weighted residual matrix are minimized. (These estimators coincide for the univariate regression model.) The common way to construct the estimator is to minimize the Frobenius norm. The estimator that minimizes the Frobenius norm also minimizes the spectral norm. Any estimator that minimizes the spectral norm is consistent under conditions of our consistency theorems (see Theorems 3.5–3.7 in Section 3). We also provide a sufficient condition for uniqueness of the estimator that minimizes the Frobenius norm.

In this paper, for the results on consistency of the TLS estimator which are stated in paper [18], we provide complete and comprehensive proofs and present all necessary auxiliary and complementary results. For convenience of the reader we first present the sketch of proof. Detailed proofs are postponed to the appendix. Moreover, the paper contains new results on the relation between the TLS estimator and the generalized eigenvalue problem.

The structure of the paper is as follows. In Section 2 we introduce the model and define the TLS estimator. The consistency theorems for different moment conditions on the errors and for different senses of consistency are stated in Section 3, and their proofs are sketched in Section 5. Section 4 states the existence and uniqueness of the TLS estimator. Auxiliary theoretical constructions and theorems are presented in Section 6. Section 7 explains the relationship between the TLS estimator and the generalized eigenvalue problem. The results in Section 7 are used in construction of the TLS estimator and in the proof of its uniqueness. Detailed proofs are moved to the appendix (Section 8).

Notations

At first, we list the *general notation*. For $v = (x_k)_{k=1}^n$ being a vector, $\|v\| = \sqrt{\sum_{k=1}^n x_k^2}$ is the 2-norm of v.

For $M = (x_{i,j})_{i=1}^m{}_{j=1}^n$ being an $m \times n$ matrix, $\|M\| = \max_{v \neq 0} \frac{\|Mv\|}{\|v\|} = \sigma_{\max}(M)$ is the spectral norm of M; $\|M\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n x_{i,j}^2}$ is the Frobenius norm of M; $\sigma_{\max}(M) = \sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_{\min(m,n)}(M) \geq 0$ are the singular values of M, arranged in descending order; $\operatorname{span}\langle M\rangle$ is the column space of M; $\operatorname{rk} M$ is the rank of M. For a square $n \times n$ matrix M, $\operatorname{def} M = n - \operatorname{rk} M$ is rank deficiency of

M; $\operatorname{tr} M = \sum_{i=1}^{n} x_{i,i}$ is the trace of M; $\chi_M(\lambda) = \det(M - \lambda I)$ is the characteristic polynomial of M. If M is an $n \times n$ matrix with real eigenvalues (e.g., if M is Hermitian or if M admits a decomposition $M = AB$, where A and B are Hermitian matrices, and either A or B is positive semidefinite), $\lambda_{\min}(M) = \lambda_1(M) \le \lambda_2(M) \le \cdots \le \lambda_n(M) = \lambda_{\max}(M)$ are eigenvalues of M arranged in ascending order.

For V_1 and V_2 being linear subspaces of $\mathbb{R}^n$ of equal dimension $\dim V_1 = \dim V_2$, $\|\sin\angle(V_1, V_2)\| = \|P_{V_1} - P_{V_2}\| = \|P_{V_1}(I - P_{V_2})\|$ is the greatest sine of the canonical angles between V_1 and V_2. See Section 6.2 for more general definitions.

Now, list *the model-specific notations*. The notations (except for the matrix Σ) come from [9]. The notations are listed here only for reference; they are introduced elsewhere in this paper – in Sections 1 and 2.

n is the number of regressors, i.e., the number of explanatory variables for each observation; d is the number of response variables for each observation; m is the number of observations, i.e., the sample size.

$C_0 = (A_0,\ B_0) = \begin{pmatrix} (a_1^0)^\top & (a_1^0)^\top X_0 \\ \cdots & \cdots \\ (a_m^0)^\top & (a_m^0)^\top X_0 \end{pmatrix} = \begin{pmatrix} (c_1^0)^\top \\ \cdots \\ (c_m^0)^\top \end{pmatrix}$ is the matrix of true variables. It is an $m \times (n+d)$ nonrandom matrix. The left-hand block A_0 of size $m \times n$ consists of true explanatory variables, and the right-hand block B_0 of size $m \times d$ consists of true response variables.

$\widetilde{C} = (\tilde{A},\ \widetilde{B}) = \begin{pmatrix} \tilde{a}_1^\top & \tilde{b}_1^\top \\ \cdots & \cdots \\ \tilde{a}_m^\top & \tilde{b}_m^\top \end{pmatrix} = \begin{pmatrix} \tilde{c}_1^\top \\ \cdots \\ \tilde{c}_m^\top \end{pmatrix} = \begin{pmatrix} \delta_{1,1} & \dots & \delta_{1,n+d} \\ \cdots & \cdots & \cdots \\ \delta_{m,1} & \dots & \delta_{m,n+d} \end{pmatrix}$ is the matrix of errors. It is an $m \times (n + d)$ random matrix.

$C = (A,\ B) = C_0 + \widetilde{C} = \begin{pmatrix} a_1^\top & b_1^\top \\ \cdots & \cdots \\ a_m^\top & b_m^\top \end{pmatrix}$ is the matrix of observations. It is an $m \times (n + d)$ random matrix.

Σ is a covariance matrix of errors for one observation. For every i, it is assumed that $\mathbb{E}\,\tilde{c}_i = 0$ and $\mathbb{E}\,\tilde{c}_i\tilde{c}_i^\top = \Sigma$. The matrix Σ is symmetric, positive semidefinite, nonrandom, and of size $(n + d) \times (n + d)$. It is assumed known when we construct the TLS estimator.

X_0 is the matrix of true regression parameters. It is a nonrandom $n \times d$ matrix and is a parameter of interest.

$X_{\mathrm{ext}}^0 = \binom{X_0}{-I}$ is an augmented matrix of regression coefficients. It is a nonrandom $(n + d) \times d$ matrix.

$\widehat{X}$ is the TLS estimator of the matrix X_0.

$\widehat{X}_{\mathrm{ext}}$ is a matrix whose column space $\operatorname{span}\langle \widehat{X}_{\mathrm{ext}} \rangle$ is considered an estimator of the subspace $\operatorname{span}\langle X_{\mathrm{ext}}^0 \rangle$. The matrix $\widehat{X}_{\mathrm{ext}}$ is of size $(n + d) \times d$. For fixed m and Σ, $\widehat{X}_{\mathrm{ext}}$ is a Borel measurable function of the matrix C.

While in consistency theorems m tends to ∞, all matrices in this list except Σ, X_0 and X_{ext}^0 silently depend on m. For example, in equations "$\lim_{m\to\infty} \lambda_{\min}(A_0^\top A_0) = +\infty$" and "$\widehat{X} \to X_0$ almost surely" the matrices A_0 and $\widehat{X}$ depend on m.

2 The model and the estimator

2.1 Statistical model

It is assumed that the matrices A_0 and B_0 satisfy the relation

$$\underset{m\times n}{A_0} \cdot \underset{n\times d}{X_0} = \underset{m\times d}{B_0}\,. \tag{2}$$

They are observed with measurement errors $\tilde{A}$ and $\widetilde{B}$, that is

$$A = A_0 + \tilde{A}, \qquad B = B_0 + \widetilde{B}.$$

The matrix X_0 is a parameter of interest.

Rewrite the relation in an implicit form. Let the $m \times (n+d)$ block matrices $C_0, \widetilde{C}, C \in \mathbb{R}^{m\times(n+d)}$ be constructed by binding "respective versions" of matrices A and B:

$$C_0 = [A_0 \; B_0], \qquad \widetilde{C} = [\tilde{A} \; \widetilde{B}], \qquad C = [A \; B].$$

Denote $X_{\mathrm{ext}}^0 = \left(\begin{smallmatrix} X_0 \\ -I_d \end{smallmatrix}\right)$. Then

$$\underset{m\times(n+d)}{C_0} \cdot \underset{(n+d)\times d}{X_{\mathrm{ext}}^0} = \underset{m\times d}{0}\,. \tag{3}$$

The entries of the matrix $\widetilde{C}$ are denoted δ_{ij}; the rows are $\tilde{c}_i$:

$$\widetilde{C} = (\delta_{ij})_{i=1}^{m}{}_{j=1}^{n+d}, \qquad \tilde{c}_i = (\delta_{ij})_{j=1}^{n+d}.$$

Throughout the paper the following three conditions are assumed to be true:

The rows $\tilde{c}_i$ of the matrix $\widetilde{C}$ are mutually independent random vectors. (4)

$\mathbb{E}\,\widetilde{C} = 0$, and $\mathbb{E}\,\tilde{c}_i \tilde{c}_i^\top := (\mathbb{E}\,\delta_{ij}\delta_{ik})_{i=1,\;k=1}^{n+d\;n+d} = \varSigma$ for all $i{=}1,\dots,m$. (5)

$\operatorname{rk}(\varSigma X_{\mathrm{ext}}^0) = d$. (6)

Example 2.1 (simple univariate linear regression with intercept)**.** For $i = 1,\dots,m$

$$\begin{cases} x_i = \xi_i + \delta_i; \\ y_i = \beta_0 + \beta_1 \xi_i + \varepsilon_i, \end{cases}$$

where the measurement errors δ_i, ε_i, $i = 1,\dots,m$, – all the $2m$ variables – are uncorrelated, $\mathbb{E}\,\delta_i = 0$, $\mathbb{E}\,\delta_i^2 = \sigma_\delta^2$, $\mathbb{E}\,\varepsilon_i = 0$, and $\mathbb{E}\,\varepsilon_i^2 = \sigma_\varepsilon^2$. A sequence $\{(x_i, y_i),\ i = 1,\dots,m\}$ is observed. The parameters β_0 and β_1 are to be estimated.

This example is taken from [1, Section 1.1]. But the notation in Example 2.1 and elsewhere in the paper is different. Our notation is $a_i^0 = (1, \xi_i)^\top$, $b_i^0 = \eta_i$, $a_i = (1, x_i)^\top$, $b_i = y_i$, $\delta_{i,1} = 0$, $\delta_{i,2} = \delta_i$, $\delta_{i,3} = \varepsilon_i$, $\varSigma = \operatorname{diag}(0, \sigma_\delta^2, \sigma_\varepsilon^2)$, and $X_0 = (\beta_0, \beta_1)^\top$.

Remark 2.1. For some matrices Σ, (6) is satisfied for any $n \times d$ matrix X_0. If the matrix Σ in nonsingular, then condition (6) is satisfied. If the errors in the explanatory variables and in the response are uncorrelated, i.e., if the matrix Σ has a block-diagonal form

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & 0 \\ 0 & \Sigma_{bb} \end{pmatrix}$$

(where $\Sigma_{aa} = \mathbb{E}\, \tilde{a}_i \tilde{a}_i^\top$ and $\Sigma_{bb} = \mathbb{E}\, \tilde{b}_i \tilde{b}_i^\top$) with nonsingular matrix Σ_{bb}, then condition (6) is satisfied. For example, in the basic mixed LS-TLS problem Σ is diagonal, Σ_{bb} is nonsingular, and so (6) holds true. If the null-space of the matrix Σ (which equals $\operatorname{span}\langle \Sigma \rangle^\perp$ because Σ is symmetric) lies inside the subspace spanned by the first n (of $n + d$) standard basis vectors, then condition (6) is also satisfied. On the other hand, if $\operatorname{rk} \Sigma < d$, then condition (6) is not satisfied.

2.2 *Total least squares (TLS) estimator*

First, find the $m \times (n + d)$ matrix Δ for which the constrained minimum is attained

$$\begin{cases} \|\Delta\, (\Sigma^{1/2})^\dagger\|_F \to \min; \\ \Delta\, (I - P_\Sigma) = 0; \\ \operatorname{rk}(C - \Delta) \le n. \end{cases} \tag{7}$$

Hereafter $\Sigma^\dagger$ is the Moore–Penrose pseudoinverse matrix of the matrix Σ, P_Σ is an orthogonal projector onto the column space of Σ, $P_\Sigma = \Sigma \Sigma^\dagger$.

Now, show that the minimum in (7) is attained. The constraint $\operatorname{rk}(C - \Delta) \le n$ is satisfied if and only if all the minors of $C - \Delta$ of order $n+1$ vanish. Thus the set of all Δ that satisfy the constraints (the constraint set) is defined by $\frac{m!(n+d)!}{(n+1)!^2(m-n-1)!(d-1)!} + 1$ algebraic equations; and so it is closed. The constraint set is nonempty *almost surely* because it contains $\widetilde{C}$. The functional $\|\Delta \Sigma^\dagger\|_F$ is a pseudonorm on $\mathbb{R}^{m\times(n+d)}$, but it is a norm on the linear subspace $\{\Delta : \Delta\, (I - \Sigma^\dagger) = 0\}$, where it induces a natural subspace topology. The constraint set is closed on the subspace (with the norm), and whenever it is nonempty (i.e., almost surely), it has a minimal-norm element.

Notice that under condition (6) the constrain set is non-empty always and not just almost surely. This follows from Proposition 7.9.

For the matrix Δ that is a solution to minimization problem (7), consider the rowspace $\operatorname{span}\langle (C - \Delta)^\top \rangle$ of the matrix $C - \Delta$. Its dimension does not exceed n. Its orthogonal basis can be completed to the orthogonal basis in $\mathbb{R}^{n+d}$, and the complement consists of $n + d - \operatorname{rk}(C - \Delta) \ge d$ vectors. Choose d vectors from the complement, which are linearly independent, and bind them (as column-vectors) into $(n + d) \times d$ matrix $\widehat{X}_{\mathrm{ext}}$. The matrix $\widehat{X}_{\mathrm{ext}}$ satisfies the equation

$$(C - \Delta)\widehat{X}_{\mathrm{ext}} = 0. \tag{8}$$

If the lower $d \times d$ block of the matrix $\widehat{X}_{\mathrm{ext}}$ is a nonsingular matrix, by linear transformation of columns (i.e., by right-multiplying by some nonsingular matrix) the matrix $\widehat{X}_{\mathrm{ext}}$ can be transformed to the form

$$\begin{pmatrix} \widehat{X} \\ -I \end{pmatrix},$$

where I is $d \times d$ identity matrix. The matrix $\widehat{X}$ satisfies the equation

$$(C - \Delta) \begin{pmatrix} \widehat{X} \\ -I \end{pmatrix} = 0. \tag{9}$$

(Otherwise, if the lower block of the matrix $\widehat{X}_{\mathrm{ext}}$ is singular, then our estimation fails. Note that whether the lower block of the matrix $\widehat{X}_{\mathrm{ext}}$ is singular might depend not only on the observations C, but also on the choice of the matrix Δ where the minimum in (7) in attained and the d vectors that make matrix $\widehat{X}_{\mathrm{ext}}$. We will show that the lower block of the matrix $\widehat{X}_{\mathrm{ext}}$ is nonsingular with high probability regardless of the choice of Δ and $\widehat{X}_{\mathrm{ext}}$.)

Columns of the matrix $\widehat{X}_{\mathrm{ext}}$ should span the eigenspace (generalized invariant space) of the matrix pencil $\langle C^\top C, \Sigma\rangle$ which corresponds to the d smallest generalized eigenvalues. That the columns of the matrix $\widehat{X}_{\mathrm{ext}}$ span the generalized invariant space corresponding to finite generalized eigenvalues is written in the matrix notation as follows:

$$\exists M \in \mathbb{R}^{d\times d} : \; C^\top C \widehat{X}_{\mathrm{ext}} = \Sigma \widehat{X}_{\mathrm{ext}} M.$$

Possible problems that may arise in the course of solving the minimization problem (7) are discussed in [18]. We should mention that our two-step definition (7) & (9) of the TLS estimator is slightly different from the conventional definition in [20, Sections 2.3.2 and 3.2] or in [10]. In these papers, the problem from which the estimator $\widehat{X}$ is found is equivalent to the following:

$$\begin{cases} \|\Delta\, (\Sigma^{1/2})^\dagger\|_F \to \min; \\ \Delta\, (I - P_\Sigma) = 0; \\ (C - \Delta) \begin{pmatrix} \widehat{X} \\ -I \end{pmatrix} = 0, \end{cases} \tag{10}$$

where the optimization is performed for Δ and $\widehat{X}$ that satisfy the constraints in (10). If our estimation defined with (7) and (9) succeeds, then the minimum values in (7) and (10) coincide, and the minimum in (10) is attained for $(\Delta, \widehat{X})$ that is the solution to (7) & (9). Conversely, if our estimation succeeds for at least one choice of Δ and $\widehat{X}_{\mathrm{ext}}$, then all the solutions to (10) can be obtained with different choices of Δ and $\widehat{X}_{\mathrm{ext}}$. However, strange things may happen if our estimation always fails.

Besides (7), consider the optimization problem

$$\begin{cases} \lambda_{\max}(\Delta \Sigma^\dagger \Delta^\top) \to \min; \\ \Delta\, (I - P_\Sigma) = 0; \\ \operatorname{rk}(C - \Delta) \le n. \end{cases} \tag{11}$$

It will be shown that every Δ that minimizes (7) also minimizes (11).

We can construct the optimization problem that generalizes both (7) and (11). Let $\|M\|_{\mathrm{U}}$ be a unitarily invariant norm on $m \times (n+d)$ matrices. Consider the optimization problem

$$\begin{cases} \|\Delta\, (\Sigma^{1/2})^\dagger\|_{\mathrm{U}} \to \min; \\ \Delta\, (I - P_\Sigma) = 0; \\ \operatorname{rk}(C - \Delta) \le n. \end{cases} \tag{12}$$

Then every Δ that minimizes (7) also minimizes (12), and every Δ that minimizes (12) also minimizes (11). If $\|M\|_{\mathrm{U}}$ is the Frobenius norm, then optimization problems (7) and (12) coincide, and if $\|M\|_{\mathrm{U}}$ is the spectral norm, then optimization problems (11) and (12) coincide.

Remark 2.2. A solution to problem (7) or (11) does not change if the matrix Σ is multiplied by a positive scalar factor. Thus, instead of assuming that the matrix Σ is known completely, we can assume that Σ is known up to a scalar factor.

3 Known consistency results

In this section we briefly revise known consistency results. One of conditions for the consistency of the TLS estimator is the convergence of $\frac{1}{m}A_0^\top A_0$ to a nonsingular matrix. It is required, for example, in [5]. The condition is relaxed in the paper by Gallo [4].

Theorem 3.1 (Gallo [4], Theorem 2). *Let $d = 1$,*

$$\begin{aligned} m^{-1/2}\lambda_{\min}\big(A_0^\top A_0\big) &\to \infty \quad as \quad m \to \infty, \\ \frac{\lambda_{\min}^2(A_0^\top A_0)}{\lambda_{\max}(A_0^\top A_0)} &\to \infty \quad as \quad m \to \infty, \end{aligned}$$

and the measurement errors $\tilde{c}_i$ are identically distributed, with finite fourth moment $\mathbb{E}\,\|\tilde{c}_i\|^4 < \infty$. Then $\widehat{X} \xrightarrow{\mathrm{P}} X_0$, $m \to \infty$.

The theorem can be generalized for the multivariate regression. The condition that the errors on different observations have the same distribution can be dropped. Instead, Kukush and Van Huffel [10] assume that the fourth moments of the error distributions are bounded.

Theorem 3.2 (Kukush and Van Huffel [10], Theorem 4a). *Let*

$$\begin{aligned} \sup_{\substack{i\ge 1\\ j=1,\dots,n+d}} \mathbb{E}\,|\delta_{ij}|^4 &< \infty, \\ m^{-1/2}\lambda_{\min}\big(A_0^\top A_0\big) &\to \infty \quad as \quad m \to \infty, \\ \frac{\lambda_{\min}^2(A_0^\top A_0)}{\lambda_{\max}(A_0^\top A_0)} &\to \infty \quad as \quad m \to \infty. \end{aligned}$$

Then $\widehat{X} \xrightarrow{\mathrm{P}} X_0$ as $m \to \infty$.

Here is the strong consistency theorem:

Theorem 3.3 (Kukush and Van Huffel [10], Theorem 4b). *Let for some $r \ge 2$ and $m_0 \ge 1$,*

$$\sup_{\substack{i\ge 1\\ j=1,\dots,n+d}} \mathbb{E}\,|\delta_{ij}|^{2r} < \infty,$$

$$\sum_{m=m_0}^{\infty}\left(\frac{\sqrt{m}}{\lambda_{\min}(A_0^\top A_0)}\right)^r < \infty,$$

$$\sum_{m=m_0}^{\infty}\left(\frac{\lambda_{\max}(A_0^\top A_0)}{\lambda_{\min}^2(A_0^\top A_0)}\right)^r < \infty.$$

Then $\widehat{X} \to X_0$ *as* $m \to \infty$*, almost surely.*

In the following consistency theorem the moment condition imposed on the errors is relaxed.

Theorem 3.4 (Kukush and Van Huffel [10], Theorem 5b)**.** *Let for some* r*,* $1 \le r < 2$*,*

$$\sup_{\substack{i\ge 1\\ j=1,\dots,n+d}} \mathbb{E}\,|\delta_{ij}|^{2r} < \infty,$$

$$m^{-1/r}\lambda_{\min}\big(A_0^\top A_0\big) \to \infty \quad \textit{as} \quad m \to \infty,$$

$$\frac{\lambda_{\min}^2(A_0^\top A_0)}{\lambda_{\max}(A_0^\top A_0)} \to \infty \quad \textit{as} \quad m \to \infty.$$

Then $\widehat{X} \xrightarrow{\mathrm{P}} X_0$ *as* $m \to \infty$*.*

Generalizations of Theorems 3.2, 3.3, and 3.4 are obtained in [18]. An essential improvement is achieved. Namely, it is not required that $\lambda_{\min}^{-2}(A_0^\top A_0)\lambda_{\max}(A_0^\top A_0)$ converge to 0.

Theorem 3.5 (Shklyar [18], Theorem 4.1, generalization of Theorems 3.2 and 3.4)**.** *Let for some* r*,* $1 \le r \le 2$*,*

$$\sup_{\substack{i\ge 1\\ j=1,\dots,n+d}} \mathbb{E}\,|\delta_{ij}|^{2r} < \infty,$$

$$m^{-1/r}\lambda_{\min}\big(A_0^\top A_0\big) \to \infty \quad \textit{as} \quad m \to \infty.$$

Then $\widehat{X} \xrightarrow{\mathrm{P}} X_0$ *as* $m \to \infty$*.*

Theorem 3.6 (Shklyar [18], Theorem 4.2, generalization of Theorem 3.3)**.** *Let for some* $r \ge 2$ *and* $m_0 \ge 1$*,*

$$\sup_{\substack{i\ge 1\\ j=1,\dots,n+d}} \mathbb{E}\,|\delta_{ij}|^{2r} < \infty,$$

$$\sum_{m=m_0}^{\infty}\left(\frac{\sqrt{m}}{\lambda_{\min}(A_0^\top A_0)}\right)^r < \infty.$$

Then $\widehat{X} \to X_0$ *as* $m \to \infty$*, almost surely.*

In the next theorem strong consistency is obtained for $r < 2$.

Theorem 3.7 (Shklyar [18], Theorem 4.3)**.** *Let for some r ($1 \le r \le 2$) and $m_0 \ge 1$,*

$$\sup_{\substack{i\ge 1\\ j=1,\dots,n+d}} \mathbb{E}\,|\delta_{ij}|^{2r} < \infty, \qquad \sum_{m=m_0}^{\infty} \frac{1}{\lambda_{\min}^{r}(A_0^\top A_0)} < \infty.$$

Then $\widehat{X} \to X_0$ as $m \to \infty$, almost surely.

The key point of the proof is the application of our own theorem on perturbation bounds for generalized eigenvectors (Theorems 6.5 and 6.6, see also [18]). The conditions were relaxed by renormalization of the data.

4 Existence and uniqueness of the estimator

When we speak of sequence $\{A_m,\ m \ge 1\}$ of random events parametrized by sample size m, we say that a random event occurs *with high probability* if the probability of the event tends to 1 as $m \to \infty$, and we say that a random event occurs *eventually* if almost surely there exists m_0 such that the random event occurs whenever $m > m_0$, that is $\mathbb{P}(\liminf_{m\to\infty} A_m) = 1$. (In this definition, A_m are random events. Elsewhere in this paper, A_m are matrices.)

Theorem 4.1. *Under the conditions of Theorem 3.5, the following three events occur with high probability; under the conditions of Theorem 3.6 or 3.7, the following relations occur eventually.*

1. *The constrained minimum in* (7) *is attained. If Δ satisfies the constraints in* (7) (*particularly, if matrix Δ is a solution to optimization problem* (7))*, then the linear equation* (8) *has a solution $\widehat{X}_{\mathrm{ext}}$ that is a full-rank matrix.*
2. *The optimization problem* (7) *has a unique solution Δ.*
3. *For any Δ that is a solution to* (7)*, equation* (9) *(which is a linear equation in $\widehat{X}$) has a unique solution.*

Theorem 4.2.

1. *The constrained minimum in* (11) *is attained. If Δ satisfies the constraints in* (11)*, then the linear equation* (8) *has a solution $\widehat{X}_{\mathrm{ext}}$ that is a full-rank matrix.*
2. *Under the conditions of Theorem 3.5, the following random event occurs with high probability: for any Δ that is a solution to* (11)*, equation* (9) *has a solution $\widehat{X}$. (Equation* (9) *might have multiple solutions.) The solution is a consistent estimator of X_0, i.e., $\widehat{X} \to X_0$ in probability.*
3. *Under the conditions of Theorem 3.6 or 3.7, the following random event occurs eventually: for any Δ that is a solution to* (11)*, equation* (9) *has a solution $\widehat{X}$. The solution is a strongly consistent estimator of X_0, i.e., $\widehat{X} \to X_0$ almost surely.*

Remark 4.2-1. Theorem 4.2 can be generalized in the following way: all references to (11) can be changed into references to (12). Thus, if Frobenius norm in the definition of the estimator is changed to any unitarily invariant norm, the consistency results are still valid.

5 Sketch of the proof of Theorems 3.5–3.7

Denote

$$N = C_0^\top C_0 + \lambda_{\min}\big(A_0^\top A_0\big) I.$$

Under the conditions of any of the consistency theorems in Section 3 there is a convergence $\lambda_{\min}(A_0^\top A_0) \to \infty$. Hence the matrix N is nonsingular for m large enough. The matrix N is used as the denominator in the law of large numbers. Also, it is used for rescaling the problem: the condition number of $N^{-1/2} C_0^\top C_0 N^{-1/2}$ equals 2 at most.

The proofs of consistency theorems differ one from another, but they have the same structure and common parts. First, the law of large numbers

$$N^{-1/2}\big(C^\top C - C_0^\top C_0 - m\Sigma\big)N^{-1/2} = N^{-1/2} \sum_{i=1}^{m} \big(c_i^\top c_i - \big(c_i^0\big)^\top c_i^0 - \Sigma\big) N^{-1/2} \to 0 \tag{13}$$

holds either in probability or almost surely, which depends on the theorem being proved. The proof varies for different theorems.

The inequalities (54) and (57) imply that whenever convergence (13) occurs, the sine between vectors $\widehat{X}_{\mathrm{ext}}$ and X_{ext}^0 (in the univariate regression) or the largest of sines of canonical values between column spans of matrices $\widehat{X}_{\mathrm{ext}}$ and X_{ext}^0 tends to 0 as the sample size m increases:

$$\big\| \sin \angle(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0)\big\| \le \big\| \sin \angle\big(N^{1/2}\widehat{X}_{\mathrm{ext}}, N^{1/2} X_{\mathrm{ext}}^0\big)\big\| \to 0. \tag{14}$$

To prove (14), we use some algebra, the fact that X_{ext}^0 (in the univariate model) or the columns of X_{ext}^0 (in the multivariate model) are the minimum-eigenvalue eigenvectors of matrix N (see ineq. (52)), and eigenvector perturbation theorems – Lemma 6.5 or Lemma 6.6.

Then, by Theorem 8.3 we conclude that

$$\|\widehat{X} - X_0\| \to 0. \tag{15}$$

6 Relevant classical results

We use some classical results. However, we state them in a form convenient for our study and provide the proof for some of them.

6.1 Generalized eigenvectors and eigenvalues

In this paper we deal with real matrices. Most theorems in this section can be generalized for matrices with complex entries by requiring that matrices be Hermitian rather than symmetric, and by complex conjugating where it is necessary.

Theorem 6.1 (Simultaneous diagonalization of a definite matrix pair)**.** *Let A and B be $n \times n$ symmetric matrices such that for some α and β the matrix $\alpha A + \beta B$ is positive definite. Then there exist a nonsingular matrix T and diagonal matrices Λ and* M *such that*

$$A = \big(T^{-1}\big)^\top \Lambda T^{-1}, \qquad B = \big(T^{-1}\big)^\top \mathrm{M} T^{-1}.$$

If in the decomposition $T = [u_1, u_2, \dots, u_n]$, $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $\mathrm{M} = \operatorname{diag}(\mu_1, \dots, \mu_n)$, then the numbers $\lambda_i/\mu_i \in \mathbb{R} \cup \{\infty\}$ are called generalized eigenvalues, and the columns u_i of the matrix T are called the right generalized eigenvectors of the matrix pencil $\langle A, B\rangle$ because the following relation holds true:

$$\mu_i A u_i = \lambda_i B u_i.$$

Theorem 6.1 is well known; see Theorem IV.3.5 in [19, page 318]. The conditions of Theorem 6.1 can be changed as follows:

Theorem 6.2. *Let A and B be symmetric positive semidefinite matrices. Then there exist a nonsingular matrix T and diagonal matrices Λ and* M *such that*

$$A = \left(T^{-1}\right)^\top \Lambda T^{-1}, \qquad B = \left(T^{-1}\right)^\top \mathrm{M} T^{-1}. \tag{16}$$

In Theorem 6.1 λ_i and μ_i cannot be equal to 0 for the same i, while in Theorem 6.2 they can. On the other hand, in Theorem 6.1 λ_i and μ_i can be any real numbers, while in Theorem 6.2 $\lambda_i \geq 0$ and $\mu_i \geq 0$. Theorem 6.2 is proved in [15].

Remark 6.2-1. If the matrices A and B are symmetric and positive semidefinite, then

$$\operatorname{rk}\langle A, B\rangle = \operatorname{rk}(A + B), \tag{17}$$

where

$$\operatorname{rk}\langle A, B\rangle = \max_k \operatorname{rk}(A + kB)$$

is the *determinantal rank* of the matrix pencil $\langle A, B\rangle$. (For square $n \times n$ matrices A and B, the determinantal rank characterizes if the matrix pencil is regular or singular. The matrix pencil $\langle A, B\rangle$ is regular if $\operatorname{rk}\langle A, B\rangle = n$, and singular if $\operatorname{rk}\langle A, B\rangle < n$.)

The inequality $\operatorname{rk}\langle A, B\rangle \geq \operatorname{rk}(A + B)$ follows from the definition of the determinantal rank. For all $k \in \mathbb{R}$ and for all such vectors x that $(A + B)x = 0$ we have $x^\top A x + x^\top B x = 0$, and because of positive semidefiniteness of matrices A and B, $x^\top A x \geq 0$ and $x^\top B x \geq 0$. Thus, $x^\top A x = x^\top B x = 0$. Again, due to positive semidefiniteness of A and B, $Ax = Bx = 0$ and $(A + kB)x = 0$. Thus, for all $k \in \mathbb{R}$

$$\begin{gathered}\left\{x : (A + B)x = 0\right\} \subset \left\{x : (A + kB)x = 0\right\},\\ \operatorname{rk}(A + B) \geq \operatorname{rk}(A + kB),\\ \operatorname{rk}\langle A, B\rangle = \max_k \operatorname{rk}(A + kB) \leq \operatorname{rk}(A + B),\end{gathered}$$

and (17) is proved.

Remark 6.2-2. Let A and B be positive semidefinite matrices of the same size such that $\operatorname{rk}(A+B) = \operatorname{rk}(B)$. The representation (16) might be not unique. But there exists a representation (16) such that

$$\begin{aligned}&\lambda_i = \mu_i = 0 \quad \text{if} \quad i = 1, \dots, \operatorname{def}(B),\\ &\mu_i > 0 \quad \text{if} \quad i = \operatorname{def}(B) + 1, \dots, n,\\ &T = \Big[\underset{n\times\operatorname{def}(B)}{T_1} \quad \underset{n\times\operatorname{rk}(B)}{T_2}\Big],\end{aligned}$$

$$T_1^\top T_2 = 0.$$

(Here if the matrix B is nonsingular, then T_1 is $n \times 0$ empty matrix; if $B = 0$, then T_2 is $n \times 0$ matrix. In these marginal cases, $T_1^\top T_2$ is an empty matrix and is considered to be zero matrix.) The desired representation can be obtained from [2] for $S = 0$ (in de Leeuw's notation). This representation is constructed as follows. Let the columns of matrix T_1 make the orthogonal normalized basis of $\operatorname{Ker}(B) = \{v : Bv = 0\}$. There exists $n \times \operatorname{rk}(B)$ matrix F such that $B = FF^\top$. Let the columns of matrix L be the orthogonal normalized eigenvectors of the matrix $F^\dagger A(F^\dagger)^\top$. Then set $T_2 = (F^\dagger)^\top L$. Note that the notation S, F and L is borrowed from [2], and is used only once. Elsewhere in the paper, the matrix F will have a different meaning.

Proposition 6.3. *Let A and B be symmetric positive semidefinite matrices such that* $\operatorname{rk}(A + B) = \operatorname{rk}(B)$. *In the simultaneous diagonalization in Theorem 6.2 with Remark 6.2-2*

$$B^\dagger = T\mathrm{M}^\dagger T^\top,$$
$$\mathrm{M}^\dagger = \operatorname{diag}\big(\underbrace{0, \ldots, 0}_{\operatorname{def}(B)}, \mu_{\operatorname{def}(B)+1}^{-1}, \ldots, \mu_n^{-1}\big).$$

Proof. Let us verify the Moore–Penrose conditions:

$$\big(T^{-1}\big)^\top \mathrm{M}T^{-1}\, T\mathrm{M}^\dagger T^\top \big(T^{-1}\big)^\top \mathrm{M}T^{-1} = \big(T^{-1}\big)^\top \mathrm{M}T^{-1}, \tag{18}$$
$$T\mathrm{M}^\dagger T^\top \big(T^{-1}\big)^\top \mathrm{M}T^{-1}\, T\mathrm{M}^\dagger T^\top = T\mathrm{M}^\dagger T^\top, \tag{19}$$

and the fact that the matrices $(T^{-1})^\top \mathrm{M}T^{-1}\, T\mathrm{M}^\dagger T^\top$ and $T\mathrm{M}^\dagger T^\top \times (T^{-1})^\top \mathrm{M}T^{-1}$ are symmetric. The equalities (18) and (19) can be verified directly; and the symmetry properties can be reduced to the equality

$$\big(T^{-1}\big)^\top P_\mathrm{M} T^\top = T P_\mathrm{M} T^{-1} \tag{20}$$

with $P_\mathrm{M} = \mathrm{M}\mathrm{M}^\dagger = \operatorname{diag}(\underbrace{0, \ldots, 0}_{\operatorname{def}(B)}, \underbrace{1, \ldots, 1}_{\operatorname{rk}(B)})$.

Since $T_1^\top T_2 = 0$, $T^\top T$ is a block diagonal matrix. Hence $P_\mathrm{M} T^\top T = T^\top T P_\mathrm{M}$, whence (20) follows. □

6.2 *Angle between two linear subspaces*

Let V_1 and V_2 be linear subspaces of $\mathbb{R}^n$, with $\dim V_1 = k_1 \le \dim V_2 = k_2$. Then there exists an orthogonal $n \times n$ matrix U such that

$$V_1 = \operatorname{span}\left\langle U \begin{pmatrix} \operatorname{diag}_{k_2 \times k_1}(\cos\theta_i,\ i = 1, \ldots, k_1) \\ \operatorname{diag}_{(n-k_2)\times k_1}(\sin\theta_i,\ i = 1, \ldots, \min(n - k_2,\ k_1)) \end{pmatrix} \right\rangle, \tag{21}$$
$$V_2 = \operatorname{span}\left\langle U \begin{pmatrix} I_{k_2} \\ 0_{(n-k_2)\times k_2} \end{pmatrix} \right\rangle. \tag{22}$$

Here rectangular diagonal matrices are allowed. If in (21) there are more cosines than sines (i.e., if $k_2 + k_1 > n$), then the excessive cosines should be equal to 1, so the columns of the bidiagonal matrix in (21) are unit vectors (which are orthogonal to each other). Here the columns of U are the vectors of some convenient “new” basis in $\mathbb{R}^n$, so U is a transitional matrix from the standard basis to “new” basis; the columns of matrix products in $\operatorname{span}\langle\cdots\rangle$ in (21) and (22) are the vectors of the bases of subspaces V_1 and V_2; the bidiagonal matrix in (21) and the diagonal matrix in (22) are the transitional matrices from “new” basis in $\mathbb{R}^n$ to the bases in V_1 and V_2, respectively.

The angles θ_k are called the canonical angles between V_1 and V_2. They can be selected so that $0 \le \theta_k \le \frac{1}{2}\pi$ (to achieve this, we might have to reverse some vectors of the bases).

Denote P_{V_1} the matrix of the orthogonal projector onto V_1. The singular values of the matrix $P_{V_1}(I - P_{V_2})$ are equal to $\sin\theta_k$ $(k = 1, \ldots, k_1)$; besides them, there is a singular value 0 of multiplicity $n - k_1$.

Denote the greatest of the sines of the canonical eigenvalues

$$\bigl\| \sin\angle(V_1, V_2)\bigr\| = \max_{k=1,\ldots,k_1} \sin\theta_k = \bigl\| P_{V_1}(I - P_{V_2})\bigr\|. \tag{23}$$

If $\dim V_1 = 1$, $V_1 = \operatorname{span}\langle v\rangle$, then

$$\sin\angle(v, V_2) = \left\| (I - P_{V_2})\frac{v}{\|v\|}\right\| = \operatorname{dist}\left(\frac{1}{\|v\|}v, V_2\right).$$

This can be generalized for $\dim V_1 \ge 1$:

$$\bigl\| \sin\angle(V_1, V_2)\bigr\| = \max_{v\in V_1\setminus\{0\}} \left\| (I - P_{V_2})\frac{v}{\|v\|}\right\|,$$

whence

$$\begin{aligned}\bigl\| \sin\angle(V_1, V_2)\bigr\|^2 &= \max_{v\in V_1\setminus\{0\}} \frac{v^\top (I - P_{V_2})v}{\|v\|^2},\\ 1 - \bigl\| \sin\angle(V_1, V_2)\bigr\|^2 &= \min_{v\in V_1\setminus\{0\}} \frac{v^\top P_{V_2} v}{\|v\|^2}.\end{aligned} \tag{24}$$

If $\dim V_1 = \dim V_2$, then $\|\sin\angle(V_1, V_2)\| = \|P_{V_1} - P_{V_2}\|$, and therefore $\|\sin\angle(V_1, V_2)\| = \|\sin\angle(V_2, V_1)\|$. Otherwise the right-hand side of (23) may change if V_1 and V_2 are swapped (particularly, if $\dim V_1 < \dim V_2$, then $\|P_{V_1}(I - P_{V_2})\|$ may or may not be equal to 1, but always $\|P_{V_2}(I - P_{V_1})\| = 1$; see the proof of Lemma 8.2 in the appendix).

We will often omit “span” in arguments of sine. Thus, for n-row matrices X_1 and X_2, $\|\sin\angle(X_1, V_2)\| = \|\sin\angle(\operatorname{span}\langle X_1\rangle, V_2)\|$ and $\|\sin\angle(X_1, X_2)\| = \|\sin\angle(\operatorname{span}\langle X_1\rangle, \operatorname{span}\langle X_2\rangle)\|$.

Lemma 6.4. *Let V_{11}, V_2 and V_{13} be three linear subspaces in $\mathbb{R}^n$, with $\dim V_{11} = d_1 < \dim V_2 = d_2 < \dim V_{13} = d_3$ and $V_{11} \subset V_{13}$. Then there exists such a linear subspace $V_{12} \subset \mathbb{R}^n$ that $V_{11} \subset V_{12} \subset V_{13}$, $\dim V_{12} = d_2$, and $\|\sin\angle(V_{12}, V_2)\| = 1$.*

Proof. Since $\dim V_{13} + \dim V_2^{\perp} = d_3 + n - d_2 > n$, there exists a vector $v \neq 0$, $v \in V_{13} \cap V_2^{\perp}$. Since $\max(d_1, 1) \leq \dim \operatorname{span}\langle V_{11}, v\rangle \leq d_1 + 1$, it holds that

$$\dim \operatorname{span}\langle V_{11}, v\rangle \leq d_2 < \dim V_{13}.$$

Therefore, there exists a d_2-dimensional subspace V_{12} such that $\operatorname{span}\langle V_{11}, v\rangle \subset V_{12} \subset V_{13}$. Then $V_{11} \subset V_{12} \subset V_{13}$ and $v \in V_{12} \cap V_2^{\perp}$. Hence $P_{V_{12}}(I - P_{V_2})v = v$, $\|P_{V_{12}}(I - P_{V_2})\| \geq 1$, and due to equation (23), $\|\sin\angle(V_{12}, V_2)\| = 1$. Thus, the subspace V_{12} has the desired properties. □

6.3 Perturbation of eigenvectors and invariant spaces

Lemma 6.5. *Let A, B, $\tilde{A}$ be symmetric matrices, $\lambda_{\min}(A) = 0$, $\lambda_2(A) > 0$ and $\lambda_{\min}(B) \geq 0$. Let $Ax_0 = 0$ and $Bx_0 \neq 0$ (so x_0 is an eigenvector of the matrix A that corresponds to the minimum eigenvalue). Let minimum of the function*

$$f(x) := \frac{x^\top (A + \tilde{A})x}{x^\top Bx}, \qquad x^\top Bx > 0,$$

be attained at the point x_*. Then

$$\sin^2\angle(x_*, x_0) \leq \frac{\|\tilde{A}\|}{\lambda_2(A)}\left(1 + \frac{\|x_0\|^2}{x_0^\top Bx_0}\,\frac{x^\top Bx}{\|x\|^2}\right).$$

Remark 6.5-1. The function $f(x)$ may or may not attain the minimum. Thus the condition $f(x_*) = \min_{x^\top Bx>0} f(x)$ sometimes cannot be satisfied. But the theorem is still true if

$$\liminf_{x\to x_*} f(x) = \inf_{x:\, x^\top Bx>0} f(x) \tag{25}$$

and $x_* \neq 0$.

Now proclaim the multivariate generalization of Lemma 6.5. We will not generalize Remark 6.5-1. Instead, we will check that the minimum is attained when we use Lemma 6.6 (see Proposition 7.10).

Lemma 6.6. *Let A, B, $\tilde{A}$ be $n \times n$ symmetric matrices, $\lambda_i(A) = 0$ for all i=1, …, d, $\lambda_{d+1}(A) > 0$, $\lambda_{\min}(B) \geq 0$. Let X_0 be $n \times d$ matrix such that $AX_0 = 0$ and the matrix $X_0^\top BX_0$ is nonsingular. Let the functional*

$$\begin{aligned} &f(X) = \lambda_{\max}\big((X^\top BX)^{-1}X^\top(A + \tilde{A})X\big) \quad \textit{if } X \in \mathbb{R}^{n\times d} \textit{ and } X^\top BX > 0,\\ &f(X) \textit{ is not defined otherwise,} \end{aligned} \tag{26}$$

attain its minimum. Then for any point X where the minimum is attained,

$$\|\sin\angle(X, X_0)\|^2 \leq \frac{\|\tilde{A}\|}{\lambda_{d+1}(A)}\big(1 + \|B\|\,\lambda_{\max}\big((X_0^\top BX_0)^{-1}X_0^\top X_0\big)\big).$$

6.4 Rosenthal inequality

In the following theorems, a random variable ξ is called *centered* if $\mathbb{E}\,\xi = 0$.

Theorem 6.7. *Let $\nu \geq 2$ be a nonrandom real number. Then there exist $\alpha \geq 0$ and $\beta \geq 0$ such that for any set of centered mutually independent random variables $\{\xi_i, i = 1, \dots, m\}$, $m \geq 1$, the following inequality holds true:*

$$\mathbb{E}\left[\left|\sum_{i=1}^{m} \xi_i\right|^{\nu}\right] \leq \alpha \sum_{i=1}^{m} \mathbb{E}\left[|\xi_i|^{\nu}\right] + \beta\left(\sum_{i=1}^{m} \mathbb{E}\,\xi_i^2\right)^{\nu/2}.$$

Theorem 6.7 is well known; see [16, Theorem 2.9, page 59].

Theorem 6.8. *Let ν be a nonrandom real number, $1 \leq \nu \leq 2$. Then there exists $\alpha \geq 0$ such that for any set of centered mutually independent random variables $\{\xi_i, i = 1, \dots, m\}$, $m \geq 1$, the inequality holds true:*

$$\mathbb{E}\left[\left|\sum_{i=1}^{m} \xi_i\right|^{\nu}\right] \leq \alpha \sum_{i=1}^{m} \mathbb{E}\left[|\xi_i|^{\nu}\right].$$

Proof. The desired inequality is trivial for $\nu = 1$. For all $1 < \nu \leq 2$ it is a consequence of the Marcinkiewicz–Zygmund inequality

$$\mathbb{E}\left[\left|\sum_{i=1}^{m} \xi_i\right|^{\nu}\right] \leq \alpha\,\mathbb{E}\left[\left(\sum_{i=1}^{m} \xi_i^2\right)^{\nu/2}\right] \leq \alpha\,\mathbb{E}\sum_{i=1}^{m} |\xi_i|^{\nu} = \alpha \sum_{i=1}^{m} \mathbb{E}\,|\xi_i|^{\nu}.$$

Here the first inequality is due to Marcinkiewicz and Zygmund [11, Theorem 13]. The second inequality follows from the fact that for $\nu \leq 2$,

$$\left(\sum_{i=1}^{m} \xi_i^2\right)^{\nu/2} \leq \sum_{i=1}^{m} |\xi_i|^{\nu}. \qquad \square$$

7 Generalized eigenvalue problem for positive semidefinite matrices

In this section we explain the relationship between the TLS estimator and the generalized eigenvalue problem. The results of this section are important for constructing the TLS estimator. Proposition 7.9 is used to state the uniqueness of the TLS estimator.

Lemma 7.1. *Let A and B be $n \times n$ symmetric positive semidefinite matrices, with simultaneous diagonalization*

$$A = \left(T^{-1}\right)^{\top} \Lambda T^{-1}, \qquad B = \left(T^{-1}\right)^{\top} \mathrm{M} T^{-1},$$

with

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n), \qquad \mathrm{M} = \operatorname{diag}(\mu_1, \dots, \mu_n)$$

(see Theorem 6.2 for its existence). For $i = 1, \dots, n$ denote

$$\nu_i = \begin{cases} \lambda_i/\mu_i & \text{if } \mu_i > 0, \\ 0 & \text{if } \lambda_i = 0, \\ +\infty & \text{if } \lambda_i > 0,\ \mu_i = 0. \end{cases}$$

Assume that $\nu_1 \leq \nu_2 \leq \cdots \leq \nu_n$. Then

$$\nu_i = \min\left\{\lambda \geq 0 \mid \text{“}\exists V,\ \dim V = i : (A - \lambda B)|_V \leq 0\text{”}\right\}, \tag{27}$$

i.e., ν_i is the smallest number $\lambda \geq 0$, such that there exists an i-dimensional subspace $V \subset \mathbb{R}^n$, such that the quadratic form $A - \lambda B$ is negative semidefinite on V.

Remark 7.1-1. $\nu_i < \infty$ if and only if

$$\exists \lambda \; \exists V, \; \dim V = i : (A - \lambda B)|_V \leq 0.$$

Remark 7.1-2. Let $\nu_i < \infty$. The minimum in (27) is attained for V being the linear span of first i columns of the matrix T (i.e., the linear span of the eigenvectors of the matrix pencil $\langle A, B \rangle$ that correspond to the i smallest generalized eigenvalues). That is

$$(A - \nu_i B)|_V \leq 0 \quad \text{for} \quad V = \operatorname{span}\bigl\langle T \bigl({}_{0_{(n-k)\times k}}^{I_k} \bigr) \bigr\rangle.$$

In Propositions 7.2–7.5 the following optimization problem is considered. For a fixed $(n+d) \times d$ matrix X find an $m \times (n+d)$ matrix Δ where the constrained minimum is attained:

$$\begin{cases} \Delta \Sigma^\dagger \Delta^\top \to \min; \\ \Delta\,(I - P_\Sigma) = 0; \\ (C - \Delta) X = 0. \end{cases} \tag{28}$$

Here the matrix X is assumed to be of full rank:

$$\operatorname{rk} X = d. \tag{29}$$

Proposition 7.2. *1. The constraints in (28) are compatible if and only if*

$$\operatorname{span}\bigl\langle X^\top C^\top \bigr\rangle \subset \operatorname{span}\bigl\langle X^\top \Sigma \bigr\rangle. \tag{30}$$

Here $\operatorname{span}\langle M \rangle$ *is a column space of the matrix M.*

2. Let the constraints in (28) be compatible. Then the least element of the partially ordered set (in the Loewner order) $\{\Delta \Sigma^\dagger \Delta^\top : \Delta\,(I - P_\Sigma) = 0$ and $(C - \Delta)X = 0\}$ is attained for $\Delta = CX(X^\top \Sigma X)^\dagger X^\top \Sigma$ and is equal to $CX(X^\top \Sigma X)^\dagger X^\top C^\top$. This means the following:

2a. For $\Delta = CX(X^\top \Sigma X)^\dagger X^\top \Sigma$, it holds that

$$\Delta\,(I - P_\Sigma) = 0, \qquad (C - \Delta)X = 0, \tag{31}$$

$$\Delta \Sigma^\dagger \Delta^\top = CX\bigl(X^\top \Sigma X\bigr)^\dagger X^\top C^\top; \tag{32}$$

2b. For any Δ which satisfies the constraints $\Delta\,(I - P_\Sigma) = 0$ and $(C - \Delta)X = 0$,

$$\Delta \Sigma^\dagger \Delta^\top \geq CX\bigl(X^\top \Sigma X\bigr)^\dagger X^\top C^\top. \tag{33}$$

Remark 7.2-1. If the constraints are compatible, the least element (and the unique minimum) is attained at a single point. Namely, the equalities

$$\Delta\,(I - P_\Sigma) = 0, \qquad (C - \Delta)X = 0,$$
$$\Delta \Sigma^\dagger \Delta^\top = CX\bigl(X^\top \Sigma X\bigr)^\dagger X^\top C^\top$$

imply $\Delta = CX(X^\top \Sigma X)^\dagger X^\top \Sigma$.

Proposition 7.3. *Let the matrix pencil $\langle C^\top C, \Sigma \rangle$ be definite and (29) hold. The constraints in (28) are compatible if and only if the matrix $X^\top \Sigma X$ is nonsingular. Then Proposition 7.2 still holds true if $(X^\top \Sigma X)^{-1}$ is substituted for $(X^\top \Sigma X)^\dagger$.*

Proposition 7.4. *Let X be an $(n+d)\times d$ matrix which satisfies* (29) *and makes the constraints in (28) compatible. Then for $k=1,2,\dots,d$,*

$$\min_{\substack{\Delta(I-P_\Sigma)=0\\(C-\Delta)X=0}} \lambda_{k+m-d}\big(\Delta\Sigma^\dagger\Delta^\top\big)$$
$$= \min\big\{\lambda\ge 0 : \text{“}\exists V\subset \operatorname{span}\langle X\rangle,\ \dim V{=}k : \big(C^\top C-\lambda\Sigma\big)|_V\le 0\text{”}\big\}. \tag{34}$$

Remark 7.4-1. In the left-hand side of (34) the minima are attained for the same $\Delta = CX(X^\top\Sigma X)^\dagger X^\top\Sigma$ for all k (the k sets where the minima are attained have non-empty intersection; we will show that the intersection comprises of a single element).

One can choose a stack of subspaces

$$V_1\subset V_2\subset\cdots\subset V_d=\operatorname{span}\langle X\rangle$$

such that V_k is the element where the minimum in the right-hand side of (34) is attained, i.e., for all $k=1,\dots,d$,

$$\dim V_k=k,\qquad V_k\subset\operatorname{span}\langle X\rangle,\qquad \big(C^\top C-\nu_k\Sigma\big)|_{V_k}\le 0,$$

with $\nu_k=\min_{\substack{\Delta(I-P_\Sigma)=0\\(C-\Delta)X=0}}\lambda_{k+m-d}(\Delta\Sigma^\dagger\Delta^\top)$.

In Propositions 7.5 to 7.9, we will use notation from simultaneous diagonalization of matrices $C^\top C$ and Σ:

$$C^\top C=\big(T^{-1}\big)^\top \Lambda T^{-1},\qquad \Sigma=\big(T^{-1}\big)^\top \mathrm{M} T^{-1}, \tag{35}$$

where

$$\Lambda=\operatorname{diag}(\lambda_1,\dots,\lambda_{n+d}),\qquad \mathrm{M}=\operatorname{diag}(\mu_1,\dots,\mu_{n+d}),$$
$$T=[u_1,u_2,\dots,u_d,\dots,u_{n+d}].$$

If Remark 6.2-2 is applicable, let the simultaneous diagonalization be constructed accordingly. For $k=1,\dots,n{+}d$ denote

$$\nu_i=\begin{cases}\lambda_k/\mu_k & \text{if } \mu_k>0,\\ 0 & \text{if } \lambda_k=0,\\ +\infty & \text{if } \lambda_k>0,\ \mu_k=0.\end{cases}$$

Let ν_k be arranged in ascending order.

Proposition 7.5. *Let X be an $(n+d)\times d$ matrix which satisfies* (29) *and makes constraints in* (28) *compatible. Then*

$$\min_{\substack{\Delta\,(I-P_\Sigma)=0\\(C-\Delta)X=0}}\lambda_{k+m-d}\big(\Delta\Sigma^\dagger\Delta^\top\big)\ge\nu_k. \tag{36}$$

If $\nu_d<\infty$, then for $X=[u_1,u_2,\dots,u_d]$ the inequality in (36) becomes an equality.

Corollary. *In the minimization problem (11), the constrained minimum is equal to*

$$\min_{\substack{\Delta\,(I-P_\Sigma)=0\\\operatorname{rk}(C-\Delta)\le n}}\lambda_{\max}\big(\Delta\Sigma^\dagger\Delta^\top\big)=\nu_d.$$

Proposition 7.6. *In the minimization problem (7) the constrained minimum is equal to*

$$\min_{\substack{\Delta(I-P_\Sigma)=0\\ \mathrm{rk}(C-\Delta)\le n}} \left\| \left(\Delta\, \Sigma^{1/2}\right)^{\dagger} \right\|_F = \sqrt{\sum_{k=1}^{d} \nu_k}.$$

Whenever the minimum in (7) is attained for some matrix Δ, the minimum in (11) is attained for the same Δ.

Proposition 7.7. *Let $\|M\|_{\mathrm{U}}$ be an arbitrary unitarily invariant norm on $m \times n$ matrices. Singular values of the matrix M are arranged in descending order and denoted $\sigma_i(M)$:*

$$\sigma_1(M) \ge \sigma_2(M) \ge \cdots \ge \sigma_{\min(m,n)}(M) \ge 0.$$

Let M_1 and M_2 be $m \times n$ matrices. Then

1. *If $\sigma_i(M_1) \le \sigma_i(M_2)$ for all $i = 1, \dots, \min(m, n)$, then $\|M_1\|_{\mathrm{U}} \le \|M_2\|_{\mathrm{U}}$.*
2. *If $\sigma_1(M_1) < \sigma_1(M_2)$ and $\sigma_i(M_1) \le \sigma_i(M_2)$ for all $i = 2, \dots, \min(m, n)$, then $\|M_1\|_{\mathrm{U}} < \|M_2\|_{\mathrm{U}}$.*

Proposition 7.8. *Consider the optimization problem* (12) *with arbitrary unitarily invariant norm $\|M\|_{\mathrm{U}}$. Then*

1. *Any minimizer Δ to the optimization problem* (7) *also minimizes* (12).
2. *Any minimizer Δ to the optimization problem* (12) *also minimizes* (11).

Proposition 7.9. *For any Δ where the minimum in (7) is attained and the corresponding solution $\widehat{X}_{\mathrm{ext}}$ of the linear equations (8) ($\widehat{X}_{\mathrm{ext}}$ is an $(n + d) \times d$ matrix of rank d), it holds that*

$$\mathrm{span}\langle u_i : \nu_i < \nu_d\rangle \subset \mathrm{span}\langle \widehat{X}_{\mathrm{ext}}\rangle \subset \mathrm{span}\langle u_i : \nu_i \le \nu_d\rangle. \tag{37}$$

Conversely, if $\nu_d < +\infty$ and the matrix $\widehat{X}_{\mathrm{ext}}$ satisfies conditions (37), then there exists a common solution Δ to the minimization problem (7) and the linear equations (8).

As a consequence, if $\nu_d < \nu_{d+1}$, then (7) and (8) unambiguously determine $\mathrm{span}\langle \widehat{X}_{\mathrm{ext}}\rangle$ of rank d.

Proposition 7.10. *Let $\langle C^\top C, \Sigma\rangle$ be a definite matrix pencil. Then for any Δ where the minimum in (11) is attained, the corresponding solution $\widehat{X}_{\mathrm{ext}}$ of the linear equations (8) (such that* $\mathrm{rk}\, \widehat{X}_{\mathrm{ext}} = d$*) is a point where the minimum of the functional*

$$X \mapsto \lambda_{\max}\left(\left(X^\top \Sigma X\right)^{-1} X^\top C^\top C X\right), \quad X\in\mathbb{R}^{(n+d)\times d}, \quad X^\top \Sigma X>0, \tag{38}$$

is attained. It is also a point where the minimum of

$$X \mapsto \lambda_{\max}\left(\left(X^\top \Sigma X\right)^{-1} X^\top \left(C^\top C - m\Sigma\right) X\right), \tag{39}$$

is attained.

The functional (39) equals the functional (38) minus m.

8 Appendix: Proofs

Detailed proofs of Theorems 3.5–3.7

8.1 Bounds for eigenvalues of some matrices used in the proof

8.1.1 Eigenvalues of the matrix $C_0^\top C_0$

The $(n+d)\times(n+d)$ matrix $C_0^\top C_0$ is symmetric and positive semidefinite. Since $C_0 X_{\mathrm{ext}}^0 = A_0 X_0 - B_0 = 0$, the matrix $C_0^\top C_0$ is rank deficient with eigenvalue 0 of multiplicity at least d. As $A_0^\top A_0$ is a $n\times n$ principal submatrix of $C_0^\top C_0$,

$$\lambda_{d+1}\big(C_0^\top C_0\big) \ge \lambda_{\min}\big(A_0^\top A_0\big) \tag{40}$$

by the Cauchy interlacing theorem (Theorem IV.4.2 from [19] used d times).

Due to inequality (40), if the matrix $A_0^\top A_0$ is nonsingular, then $\lambda_{n+1}(C_0^\top C_0) > 0$, whence $\operatorname{rk}(C_0^\top C_0) = d$. If the conditions of Theorem 3.5, 3.6 or 3.7 hold true, then $\lambda_{\min}(A_0^\top A_0) \to \infty$, and thus

$$\lambda_{d+1}\big(C_0^\top C_0\big) \ge \lambda_{\min}\big(A_0^\top A_0\big) > 0$$

for m large enough.

Proposition 8.1. *If conditions* (4)–(6) *hold true, and conditions of either of Theorems 3.5, 3.6, or 3.7 hold true, then for m large enough $\langle C^\top C, \Sigma\rangle$ is a definite matrix pencil almost surely. More specifically,*

$$\exists m_0\ \forall m > m_0 :\ \mathbb{P}\big(C^\top C + \Sigma > 0\big) = 1.$$

Proof. *1.* If the matrix Σ is nonsingular, then Proposition 8.1 is obvious. Due to condition (6), $\operatorname{rk}\Sigma \ge d$ (see Remark 2.1), whence $\Sigma \ne 0$. In what follows, assume that Σ is a singular but non-zero matrix. Let $F = \big(\begin{smallmatrix} F_1 \\ F_2 \end{smallmatrix}\big)$ be a $(n+d)\times(n+d-\operatorname{rk}(\Sigma))$ matrix whose columns make the basis of the null-space $\operatorname{Ker}(\Sigma) = \{x : \Sigma x = 0\}$ of the matrix Σ.

2. Now prove that columns of the matrix $[I_n\ \ X_0]F$ are linearly independent. Assume the contrary. Then for some $v \in \mathbb{R}^{n+d-\operatorname{rk}(\Sigma)} \setminus \{0\}$,

$$\begin{aligned} [I_n \quad X_0]\, Fv &= 0,\\ F_1 v &= -X_0 F_2 v,\\ Fv &= \big(\begin{smallmatrix} X_0 \\ -I_d \end{smallmatrix}\big) F_2 v = X_{\mathrm{ext}}^0 F_2 v, && (41)\\ 0 = \Sigma F v &= \Sigma X_{\mathrm{ext}}^0 \cdot F_2 v. && (42) \end{aligned}$$

Furthermore, $Fv \ne 0$ because $v \ne 0$ and the columns of F are linearly independent. Hence, by (41), $F_2 v \ne 0$.

Equality (42) implies that the columns of the matrix $\Sigma X_{\mathrm{ext}}^0$ are linearly dependent, and this contradicts condition (6). The contradiction means that columns of the matrix $[I\ \ X_{\mathrm{ext}}^0]\ F$ are linearly independent.

3. If the conditions of either Theorem 3.5, 3.6, or 3.7 hold true, then the matrix $A_0^\top A_0$ is positive definite for m large enough.

4. Under conditions (4) and (5), $\tilde{C}F = 0$ almost surely. Indeed, $\mathbb{E}\,\tilde{c}_i = 0$ and $\operatorname{var}[\tilde{c}_i F] = F^\top \Sigma F = 0$, $i{=}1, 2, \dots, m$.

5. It remains to prove the implication:

$$\text{if} \quad A_0^\top A_0 > 0 \quad \text{and} \quad \tilde{C}F = 0, \quad \text{then} \quad C^\top C + \Sigma > 0.$$

The matrices $C^\top C$ and Σ are positive semidefinite. Suppose that $x^\top(C^\top C{+}\Sigma)x = 0$ and prove that $x = 0$. Since $x^\top(C^\top C + \Sigma)x = 0$, $Cx = 0$ and $\Sigma x = 0$. The vector x belongs to the null-space of the matrix Σ. Therefore, $x = Fv$ for some vector $v \in \mathbb{R}^{n+d-\operatorname{rk}\Sigma}$. Then

$$\begin{aligned} 0 = A_0^\top Cx &= A_0(C_0 + \tilde{C})x \\ &= A_0 C_0 F v + A_0 \tilde{C} F v \\ &= A_0^\top A_0 \, [I_n \quad X_0] \, F v + 0. \end{aligned} \tag{43}$$

As the matrix $A_0^\top A_0$ is nonsingular and columns of the matrix $[I_n \;\, X_0]\, F$ are linearly independent, the columns of the matrix $A_0^\top A_0 \, [I_n \;\, X_0]\, F$ are linearly independent as well. Hence, (43) implies $v = 0$, and so $x = Fv = 0$.

We have proved that the equality $x^\top(C^\top C + \Sigma)x = 0$ implies $x = 0$. Thus, the positive semidefinite matrix $C^\top C + \Sigma$ is nonsingular, and so positive definite. □

8.1.2 Eigenvalues and common eigenvectors of N and $N^{-\frac{1}{2}} C_0^\top C_0 N^{-\frac{1}{2}}$

The rank-deficient positive semidefinite symmetric matrix $C_0^\top C_0$ can be factorized as:

$$\begin{aligned} C_0^\top C_0 &= U \operatorname{diag}\bigl(\lambda_{\min}(C_0^\top C_0), \lambda_2(C_0^\top C_0), \dots, \lambda_{n+d}(C_0^\top C_0)\bigr) U^\top \\ &= U \operatorname{diag}\bigl(\lambda_j(C_0^\top C_0);\;\; j = 1, \dots, n+d\bigr) U^\top, \end{aligned}$$

with an orthogonal matrix U and

$$\lambda_{\min}\bigl(C_0^\top C_0\bigr) = \lambda_2\bigl(C_0^\top C_0\bigr) = \dots = \lambda_d\bigl(C_0^\top C_0\bigr) = 0.$$

Then the eigendecomposition of the matrix $N = C_0^\top C_0 + \lambda_{\min}(A_0^\top A_0) I$ is

$$N = U \operatorname{diag}\bigl(\lambda_j\bigl(C_0^\top C_0\bigr) + \lambda_{\min}\bigl(A_0^\top A_0\bigr);\;\; j = 1, \dots, n+d\bigr) U^\top.$$

Notice that

$$\lambda_{\min}(N) = \dots = \lambda_d(N) = \lambda_{\min}\bigl(A_0^\top A_0\bigr). \tag{44}$$

The matrix N is nonsingular as soon as $A_0^\top A_0$ is nonsingular. Hence, under the conditions of Theorem 3.5, 3.6, or 3.7, the matrix N is nonsingular for m large enough.

Since $C_0 X_{\text{ext}}^0 = 0$, it holds that

$$N X_{\text{ext}}^0 = \lambda_{\min}\bigl(A_0^\top A_0\bigr) X_{\text{ext}}^0. \tag{45}$$

As soon as N is nonsingular, the matrices $N^{-1/2}$ and $N^{-1/2}C_0^\top C_0 N^{-1/2}$ have the eigendecomposition

$$N^{-1/2} = U\,\mathrm{diag}\biggl(\frac{1}{\sqrt{\lambda_j(C_0^\top C_0)+\lambda_{\min}(A_0^\top A_0)}};\ j=1,\dots,n+d\biggr)U^\top,$$

$$N^{-1/2}C_0^\top C_0 N^{-1/2} = U\,\mathrm{diag}\biggl(\frac{\lambda_j(C_0^\top C_0)}{\lambda_j(C_0^\top C_0)+\lambda_{\min}(A_0^\top A_0)};\ j=1,\dots,n{+}d\biggr)U^\top.$$

Thus, the eigenvalues of $N^{-1/2}$ and $N^{-1/2}C_0^\top C_0 N^{-1/2}$ satisfy the following:

$$\bigl\|N^{-1/2}\bigr\| = \lambda_{\max}\bigl(N^{-1/2}\bigr) = \frac{1}{\sqrt{\lambda_{\min}(A_0^\top A_0)}}; \tag{46}$$

$$\lambda_j\bigl(N^{-1/2}C_0^\top C_0 N^{-1/2}\bigr) = 0, \quad j=1,\dots,d; \tag{47}$$

$$\tfrac{1}{2} \le \lambda_j\bigl(N^{-1/2}C_0^\top C_0 N^{-1/2}\bigr) \le 1, \quad j=d{+}1,\dots,n{+}d. \tag{48}$$

As a result,

$$\tfrac{1}{2}n \le \mathrm{tr}\bigl(N^{-1/2}C_0^\top C_0 N^{-1/2}\bigr) \le n. \tag{49}$$

Because $\mathrm{tr}(C_0 N^{-1} C_0^\top) = \mathrm{tr}(C_0 N^{-1/2} N^{-1/2} C_0^\top) = \mathrm{tr}(N^{-1/2} C_0^\top C_0 N^{-1/2})$,

$$\tfrac{1}{2}n \le \mathrm{tr}\bigl(C_0 N^{-1} C_0^\top\bigr) \le n. \tag{50}$$

These properties will be used in Sections 8.2 and 8.3.

8.2 Use of eigenvector perturbation theorems

8.2.1 Univariate regression ($d = 1$)

Remember inequalities (44) (whence (51) follows) and (45):

$$\widehat{X}_{\mathrm{ext}}^\top N \widehat{X}_{\mathrm{ext}} \ge \lambda_{\min}\bigl(A_0^\top A_0\bigr)\widehat{X}_{\mathrm{ext}}^\top \widehat{X}_{\mathrm{ext}}; \tag{51}$$

$$N X_{\mathrm{ext}}^0 = \lambda_{\min}\bigl(A_0^\top A_0\bigr) X_{\mathrm{ext}}^0.$$

Then

$$\frac{(\widehat{X}_{\mathrm{ext}}^\top X_{\mathrm{ext}}^0)^2}{\widehat{X}_{\mathrm{ext}}^\top \widehat{X}_{\mathrm{ext}} \cdot X_{\mathrm{ext}}^{0\top} X_{\mathrm{ext}}^0} \ge \frac{(\widehat{X}_{\mathrm{ext}}^\top N X_{\mathrm{ext}}^0)^2}{\widehat{X}_{\mathrm{ext}}^\top N \widehat{X}_{\mathrm{ext}} \cdot X_{\mathrm{ext}}^{0\top} N X_{\mathrm{ext}}^0},$$

$$\cos^2\angle\bigl(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0\bigr) \ge \cos^2\angle\bigl(N^{1/2}\widehat{X}_{\mathrm{ext}}, N^{1/2}X_{\mathrm{ext}}^0\bigr),$$

$$\sin^2\angle\bigl(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0\bigr) \le \sin^2\angle\bigl(N^{1/2}\widehat{X}_{\mathrm{ext}}, N^{1/2}X_{\mathrm{ext}}^0\bigr). \tag{52}$$

Now, apply Lemma 6.5 on the perturbation bound for the minimum-eigenvalue eigenvector. The unperturbed symmetric matrix is $N^{-1/2}C_0^\top C_0 N^{-1/2}$, satisfying

$$\lambda_{\min}\bigl(N^{-1/2}C_0^\top C_0 N^{-1/2}\bigr) = 0,$$

$$N^{-1/2}C_0^\top C_0 N^{-1/2} N^{1/2} X_{\mathrm{ext}}^0 = 0,$$

$$\lambda_2\bigl(N^{-1/2}C_0^\top C_0 N^{-1/2}\bigr) \ge \tfrac{1}{2}.$$

The null-vector of the unperturbed matrix is $N^{-1/2}X_{\mathrm{ext}}^0$.

The column vector $\widehat{X}_{\mathrm{ext}}$ is a generalized eigenvector of the matrix pencil $\langle C^\top C, \Sigma\rangle$. Denote the corresponding eigenvalue by $\lambda_{\min}$. Thus,

$$C^\top C \widehat{X}_{\mathrm{ext}} = \lambda_{\min} \cdot \Sigma \widehat{X}_{\mathrm{ext}}.$$

The perturbed matrix is $N^{-1/2}(C^\top C - m\Sigma)N^{-1/2}$; the minimum eigenvalue of the matrix pencil $\langle N^{-1/2}(C^\top C - m\Sigma)N^{-1/2},\ N^{-1/2}\Sigma N^{-1/2}\rangle$ is equal to $\lambda_{\min} - m$, and the eigenvector is $N^{1/2}\widehat{X}_{\mathrm{ext}}$:

$$N^{-1/2}\big(C^\top C - m\Sigma\big)N^{-1/2}N^{1/2}\widehat{X}_{\mathrm{ext}} = (\lambda_{\min} - m)N^{-1/2}\Sigma N^{-1/2}N^{1/2}\widehat{X}_{\mathrm{ext}}.$$

We have to verify that $N^{-1/2}\Sigma N^{-1/2}N^{1/2}X^0_{\mathrm{ext}} \neq 0$; this follows from condition (6). Obviously, the matrix $N^{-1/2}\Sigma N^{-1/2}$ is positive semidefinite:

$$N^{-1/2}\Sigma N^{-1/2} \geq 0. \tag{53}$$

Denote

$$\epsilon = \big\| N^{-1/2}\big(C^\top C - m\Sigma\big)N^{-1/2} - N^{-1/2}C_0^\top C_0 N^{-1/2}\big\|.$$

By Lemma 6.5

$$\sin^2 \angle\big(N^{1/2}\widehat{X}_{\mathrm{ext}}, N^{1/2}X^0_{\mathrm{ext}}\big) \leq \frac{\epsilon}{0.5}\left(1 + \frac{X^{0\top}_{\mathrm{ext}} N X^0_{\mathrm{ext}}}{X^{0\top}_{\mathrm{ext}}\Sigma X^0_{\mathrm{ext}}} \cdot \frac{\widehat{X}^\top_{\mathrm{ext}}\Sigma\widehat{X}_{\mathrm{ext}}}{\widehat{X}^\top_{\mathrm{ext}} N \widehat{X}_{\mathrm{ext}}}\right).$$

Use (45) and (51) again, and also use (52):

$$\begin{aligned}\sin^2 \angle\big(\widehat{X}_{\mathrm{ext}}, X^0_{\mathrm{ext}}\big) &\leq \sin^2 \angle\big(N^{1/2}\widehat{X}_{\mathrm{ext}}, N^{1/2}X^0_{\mathrm{ext}}\big)\\ &\leq 2\epsilon\left(1 + \frac{X^{0\top}_{\mathrm{ext}} X^0_{\mathrm{ext}}}{X^{0\top}_{\mathrm{ext}}\Sigma X^0_{\mathrm{ext}}} \cdot \frac{\widehat{X}^\top_{\mathrm{ext}}\Sigma\widehat{X}_{\mathrm{ext}}}{\widehat{X}^\top_{\mathrm{ext}} \widehat{X}_{\mathrm{ext}}}\right)\\ &\leq 2\epsilon\left(1 + \frac{X^{0\top}_{\mathrm{ext}} X^0_{\mathrm{ext}} \cdot \|\Sigma\|}{X^{0\top}_{\mathrm{ext}}\Sigma X^0_{\mathrm{ext}}}\right).\end{aligned} \tag{54}$$

8.2.2 Multivariate regression ($d \geq 1$)

What follows is valid for both univariate ($d = 1$) and multivariate ($d > 1$) regression.

Due to (44), $N \geq \lambda_{\min}(A_0^\top A_0)I$ in the Loewner order; thus inequality (51) holds in the Loewner order. Hence

$$\begin{aligned}\forall v \in \mathbb{R}^d \setminus \{0\} : &\frac{v^\top \widehat{X}^\top_{\mathrm{ext}} X^0_{\mathrm{ext}}(X^{0\top}_{\mathrm{ext}} X^0_{\mathrm{ext}})^{-1} X^{0\top}_{\mathrm{ext}} \widehat{X}_{\mathrm{ext}} v}{v^\top \widehat{X}^\top_{\mathrm{ext}} \widehat{X}_{\mathrm{ext}} v}\\ &\qquad \geq \lambda_{\min}\big(A_0^\top A_0\big)\frac{v^\top \widehat{X}^\top_{\mathrm{ext}} X^0_{\mathrm{ext}}(X^{0\top}_{\mathrm{ext}} X^0_{\mathrm{ext}})^{-1} X^{0\top}_{\mathrm{ext}} \widehat{X}_{\mathrm{ext}} v}{v^\top \widehat{X}^\top_{\mathrm{ext}} N \widehat{X}_{\mathrm{ext}} v}.\end{aligned}$$

With inequality (45), we get

$$\frac{v^\top \widehat{X}^\top_{\mathrm{ext}} X^0_{\mathrm{ext}}(X^{0\top}_{\mathrm{ext}} X^0_{\mathrm{ext}})^{-1} X^{0\top}_{\mathrm{ext}} \widehat{X}_{\mathrm{ext}} v}{v^\top \widehat{X}^\top_{\mathrm{ext}} \widehat{X}_{\mathrm{ext}} v}$$

$$\geq \frac{v^\top N \widehat{X}_{\mathrm{ext}}^\top X_{\mathrm{ext}}^0 (X_{\mathrm{ext}}^{0\top} N X_{\mathrm{ext}}^0)^{-1} X_{\mathrm{ext}}^{0\top} N \widehat{X}_{\mathrm{ext}} v}{v^\top \widehat{X}_{\mathrm{ext}}^\top N \widehat{X}_{\mathrm{ext}} v}.$$

Using equation (24) to determine the sine and noticing that

$$\begin{aligned} P_{X_{\mathrm{ext}}^0} &= X_{\mathrm{ext}}^0 \big(X_{\mathrm{ext}}^{0\top} X_{\mathrm{ext}}^0\big)^{-1} X_{\mathrm{ext}}^{0\top}, \\ P_{N^{1/2} X_{\mathrm{ext}}^0} &= N^{1/2} X_{\mathrm{ext}}^0 \big(X_{\mathrm{ext}}^{0\top} N X_{\mathrm{ext}}^0\big)^{-1} X_{\mathrm{ext}}^{0\top} N^{1/2}, \end{aligned}$$

we get

$$\begin{gathered} 1 - \big\| \sin \angle\big(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0\big) \big\|^2 \geq 1 - \big\| \sin \angle\big(N^{1/2} \widehat{X}_{\mathrm{ext}}, N^{1/2} X_{\mathrm{ext}}^0\big) \big\|^2, \\ \big\| \sin \angle\big(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0\big) \big\| \leq \big\| \sin \angle\big(N^{1/2} \widehat{X}_{\mathrm{ext}}, N^{1/2} X_{\mathrm{ext}}^0\big) \big\|. \end{gathered} \tag{55}$$

The TLS estimator $\widehat{X}_{\mathrm{ext}}$ is defined as a solution to the linear equations (8) for Δ that brings the minimum to (7). By Proposition 7.6, the same Δ brings the minimum to (11). By Proposition 7.10, the functions (38) and (39) attain their minima at the point $\widehat{X}_{\mathrm{ext}}$. Therefore, the minimum of the function

$$M \mapsto \lambda_{\max}\big(\big(M^\top N^{-1/2} \Sigma N^{-1/2} M\big)^{-1} M^\top N^{-1/2} \big(C^\top C - m\Sigma\big) N^{-1/2} M\big) \tag{56}$$

is attained for $M = N^{1/2} \widehat{X}_{\mathrm{ext}}$.

Now, apply Lemma 6.6 on perturbation bounds for a generalized invariant subspace. The unperturbed matrix (denoted A in Lemma 6.6) is $N^{-1/2} C_0^\top C_0 N^{-1/2}$; its nullspace is the column space of the matrix $N^{1/2} X_{\mathrm{ext}}^0$ (which is denoted X_0 in Lemma 6.6). The perturbed matrix ($A + \tilde{A}$ in Lemma 6.6) is $N^{-1/2}(C^\top C - m\Sigma) N^{-1/2}$. The matrix B in Lemma 6.6 equals $N^{-1/2} \Sigma N^{-1/2}$. The norm of the perturbation is denoted ϵ (it is $\|\tilde{A}\|$ in Lemma 6.6). The $(n+d) \times d$ matrix which brings the minimum to (56) is $N^{1/2} \widehat{X}_{\mathrm{ext}}$. The other conditions of Lemma 6.6 are (47), (48), and (53). We have

$$\begin{aligned} &\big\| \sin \angle\big(N^{1/2} \widehat{X}_{\mathrm{ext}}, N^{1/2} X_{\mathrm{ext}}^0\big) \big\|^2 \\ &\quad \leq \frac{\epsilon}{0.5} \big(1 + \big\|N^{-1/2} \Sigma N^{-1/2}\big\| \, \lambda_{\max}\big(\big(X_{\mathrm{ext}}^{0\top} \Sigma X_{\mathrm{ext}}^0\big)^{-1} X_{\mathrm{ext}}^{0\top} N X_{\mathrm{ext}}^0\big)\big). \end{aligned}$$

Again, with (55), (45) and (46), we have

$$\begin{aligned} &\big\| \sin \angle\big(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0\big) \big\|^2 \\ &\quad \leq \big\| \sin \angle\big(N^{1/2} \widehat{X}_{\mathrm{ext}}, N^{1/2} X_{\mathrm{ext}}^0\big) \big\|^2 \\ &\quad \leq 2\epsilon \left(1 + \frac{\|\Sigma\|}{\lambda_{\min}(A_0^\top A_0)} \lambda_{\max}\big(\lambda_{\min}\big(A_0^\top A_0\big) \big(X_{\mathrm{ext}}^{0\top} \Sigma X_{\mathrm{ext}}^0\big)^{-1} X_{\mathrm{ext}}^{0\top} X_{\mathrm{ext}}^0\big)\right) \\ &\quad = 2\epsilon \big(1 + \|\Sigma\| \, \lambda_{\max}\big(\big(X_{\mathrm{ext}}^{0\top} \Sigma X_{\mathrm{ext}}^0\big)^{-1} X_{\mathrm{ext}}^{0\top} X_{\mathrm{ext}}^0\big)\big). \end{aligned} \tag{57}$$

8.3 *Proof of the convergence* $\epsilon \to 0$

In this section, we prove the convergences

$$M_1 = N^{-1/2} C_0^\top \widetilde{C} N^{-1/2} \to 0,$$

$$M_2 = N^{-1/2}\big(\widetilde{C}^\top \widetilde{C} - m\Sigma\big)N^{-1/2} \to 0$$

in probability for Theorem 3.5, and almost surely for Theorems 3.6 and 3.7. As $\epsilon = \|M_1 + M_1^\top + M_2\|$, the convergences $M_1 \to 0$ and $M_2 \to 0$ imply $\epsilon \to 0$.

End of the proof of Theorem 3.5. It holds that

$$\begin{aligned}\|M_1\|_F^2 &= \big\|N^{-1/2}C_0^\top \tilde{C} N^{-1/2}\big\|_F^2 = \operatorname{tr}\big(N^{-1/2}C_0^\top \tilde{C} N^{-1} C_0 \tilde{C}^\top N^{-1/2}\big) \\ &= \operatorname{tr}\big(C_0 N^{-1} C_0^\top \tilde{C} N^{-1} \tilde{C}^\top\big) = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i^0 N^{-1}\big(c_j^0\big)^\top \tilde{c}_j N^{-1} \tilde{c}_i^\top .\end{aligned}$$

The right-hand side can be simplified since $\mathbb{E}\, \tilde{c}_j N^{-1} \tilde{c}_i^\top = 0$ for $i \neq j$ and $\mathbb{E}\, \tilde{c}_i N^{-1} \tilde{c}_i^\top = \operatorname{tr}(\Sigma N^{-1})$:

$$\mathbb{E}\, \|M_1\|_F^2 = \sum_{i=1}^{m} c_{0i} N^{-1} c_{0i}^\top \operatorname{tr}\big(\Sigma N^{-1}\big) = \operatorname{tr}\big(C_0 N^{-1} C_0^\top\big) \operatorname{tr}\big(\Sigma N^{-1}\big).$$

The first multiplier in the right-hand side is bounded due to (50) as $\operatorname{tr}(C_0 N^{-1} C_0^\top) \le n$, for m large enough. Now, construct an upper bound for the second multiplier:

$$\begin{aligned}\operatorname{tr}\big(\Sigma N^{-1}\big) &= \big\|N^{-1/2}\Sigma^{1/2}\big\|_F^2 \le \big\|N^{-1/2}\big\|^2 \big\|\Sigma^{1/2}\big\|_F^2 = \lambda_{\max}\big(N^{-1}\big) \operatorname{tr}\Sigma \\ &= \frac{\operatorname{tr}\Sigma}{\lambda_{\min}(N)} = \frac{\operatorname{tr}\Sigma}{\lambda_{\min}(A_0^\top A_0)}.\end{aligned}$$

Finally,

$$\mathbb{E}\, \|M_1\|_F^2 \le \frac{n \operatorname{tr}\Sigma}{\lambda_{\min}(A_0^\top A_0)}.$$

The conditions of Theorem 3.5 imply that $\lambda_{\max}(A_0^\top A_0) \to \infty$; therefore, $M_1 \xrightarrow{\mathrm{P}} 0$ as $m \to \infty$.

Now, we prove that $M_2 \xrightarrow{\mathrm{P}} 0$ as $m \to \infty$. We have

$$\begin{aligned} & M_2 = N^{-1/2}\big(\tilde{C}^\top \tilde{C} - m\Sigma\big)N^{-1/2}, \\ & \|M_2\| \le \big\|N^{-1/2}\big\| \, \big\|\tilde{C}^\top \tilde{C} - m\Sigma\big\| \, \big\|N^{-1/2}\big\| = \frac{\| \sum_{i=1}^{m} (\tilde{c}_i^\top \tilde{c}_i - \Sigma)\|}{\lambda_{\min}(A_0^\top A_0)}. \end{aligned} \tag{58}$$

Now apply the Rosenthal inequality (case $1 \le \nu \le 2$; Theorem 6.8) to construct a bound for $\mathbb{E}\, \|M_2\|^r$:

$$\mathbb{E}\, \|M_2\|^r \le \frac{\text{const} \sum_{i=1}^{m} \mathbb{E}\, \|\tilde{c}_i^\top \tilde{c}_i - \Sigma\|^r}{\lambda_{\min}^r (A_0^\top A_0)}.$$

By the conditions of Theorem 3.5, the sequence $\{\mathbb{E}\, \|\tilde{c}_i^\top \tilde{c}_i - \Sigma\|^r,\ i = 1, 2, \ldots\}$ is bounded. Hence

$$\begin{aligned} & \mathbb{E}\, \|M_2\|^r \le \frac{O(m)}{\lambda_{\min}^r (A_0^\top A_0)} \quad \text{as } m \to \infty, \\ & \mathbb{E}\, \|M_2\|^r \to 0 \quad \text{and} \quad M_2 \xrightarrow{\mathrm{P}} 0 \quad \text{as } m \to \infty. \end{aligned}$$

□

End of the proof of Theorem 3.6.

$$M_1 = \sum_{i=1}^{m} N^{-1/2} c_{0i}^\top \tilde{c}_i N^{-1/2}.$$

By the Rosenthal inequality (case $\nu \geq 2$; Theorem 6.7)

$$\mathbb{E}\,\|M_1\|^{2r} \leq \text{const} \sum_{i=1}^{m} \mathbb{E}\left\| N^{-1/2} c_{0i}^\top \tilde{c}_i N^{-1/2} \right\|^{2r} + \\ + \text{const}\left(\sum_{i=1}^{m} \mathbb{E}\,\| N^{-1/2} c_{0i}^\top \tilde{c}_i N^{-1/2} \|^2 \right)^r.$$

Construct an upper bound for the first summand:

$$\sum_{i=1}^{m} \mathbb{E}\left\| N^{-1/2} c_{0i}^\top \tilde{c}_i N^{-1/2} \right\|^{2r} \leq \sum_{i=1}^{m} \left\| N^{-1/2} c_{0i}^\top \right\|^{2r} \max_{i=1,\ldots,m} \mathbb{E}\,\|\tilde{c}_i\|^{2r} \left\| N^{-1/2} \right\|^{2r},$$

$$\sum_{i=1}^{m} \| N^{-1/2} c_{0i}^\top \|^{2r} \leq \left(\sum_{i=1}^{m} \| N^{-1/2} c_{0i}^\top \|^2 \right)^r \\ = \left(\sum_{i=1}^{m} c_{0i} N^{-1} c_{0i}^\top \right)^r = \left(\operatorname{tr}\left(C_0 N^{-1} C_0^\top \right) \right)^r \leq n^r$$

by inequality (50). By the conditions of Theorem 3.6, the sequence $\{ \max_{i=1,\ldots,m} \mathbb{E}\,\|\tilde{c}_i\|^{2r}$, $m = 1, 2, \ldots \}$ is bounded. Remember that $\|N^{-1/2}\| = \lambda_{\min}^{-1/2}(A_0^\top A_0)$. Thus,

$$\sum_{i=1}^{m} \mathbb{E}\left\| N^{-1/2} c_{0i}^\top \tilde{c}_i N^{-1/2} \right\|^{2r} = \frac{O(1)}{\lambda_{\min}^r (A_0^\top A_0)} \quad \text{as } m \to \infty.$$

The asymptotic relation

$$\sum_{i=1}^{m} \mathbb{E}\left\| N^{-1/2} c_{0i}^\top \tilde{c}_i N^{-1/2} \right\|^2 = \frac{O(1)}{\lambda_{\min} (A_0^\top A_0)}$$

can be proved similarly; in order to prove it, we use boundedness of the sequence $\{ \max_{i=1,\ldots,m} \mathbb{E}\,\|\tilde{c}_i\|^2,\ m = 1, 2, \ldots \}$. Finally,

$$\mathbb{E}\,\|M_1\|^{2r} = \frac{O(1)}{\lambda_{\min}^r (A_0^\top A_0)} \quad \text{as } m \to \infty.$$

The conditions of Theorem 3.6 imply that $\sum_{m=m_0}^{\infty} \mathbb{E}\,\|M_1\|^{2r} < \infty$, whence $M_1 \to 0$ as $m \to \infty$, almost surely.

Now, prove that $M_2 \to 0$ almost surely. In order to construct a bound for $\mathbb{E}\,\|M_2\|^r$, use the Rosenthal inequality (case $\nu \geq 2$; Theorem 6.7) as well as (58):

$$\mathbb{E}\,\|M_2\|^r \leq \frac{\mathbb{E}\,\| \sum_{i=1}^{m} (c_i^\top \tilde{c}_i - \Sigma) \|^r}{\lambda_{\min}^r (A_0^\top A_0)}$$

$$\leq \frac{\mathrm{const}\sum_{i=1}^m \mathbb{E}\,\|\tilde{c}_i^\top \tilde{c}_i - \Sigma\|^r}{\lambda_{\min}^r(A_0^\top A_0)} + \frac{\mathrm{const}(\sum_{i=1}^m \mathbb{E}\,\|\tilde{c}_i^\top \tilde{c}_i - \Sigma\|^2)^{r/2}}{\lambda_{\min}^r(A_0^\top A_0)}.$$

Under the conditions of Theorem 3.6, the sequences $\{\mathbb{E}\,\|\tilde{c}_i^\top \tilde{c}_i - \Sigma\|^r,\ i = 1, 2, \ldots\}$ and $\{\mathbb{E}\,\|\tilde{c}_i^\top \tilde{c}_i - \Sigma\|^2,\ i = 1, 2, \ldots\}$ are bounded. Thus,

$$\mathbb{E}\,\|M_2\|^r = \frac{O(m^{r/2})}{\lambda_{\min}^r(A_0^\top A_0)} \quad \text{as } m \to \infty;$$

$$\sum_{m=m_0}^{\infty} \mathbb{E}\,\|M_2\|^r < \infty,$$

whence $M_2 \to 0$ as $m \to \infty$, almost surely. □

End of the proof of Theorem 3.7. The proof of the asymptotic relation

$$\mathbb{E}\,\|M_1\|^{2r} = \frac{O(1)}{\lambda_{\min}^r(A_0^\top A_0)} \quad \text{as } m \to \infty$$

from Theorem 3.6 is still valid. The almost sure convergence $M_1 \to 0$ as $m \to \infty$ is proved in the same way as in Theorem 3.6.

Now, show that $M_2 \to 0$ as $m \to \infty$, almost surely. Under the condition of Theorem 3.7,

$$\mathbb{E}\,\big\|\tilde{c}_m^\top \tilde{c}_m - \Sigma\big\|^r = O(1), \qquad \sum_{m=m_0}^{\infty} \frac{\mathbb{E}\,\|\tilde{c}_m^\top \tilde{c}_m - \Sigma\|^r}{\lambda_{\min}^r(A_0^\top A_0)} < \infty,$$

and $\mathbb{E}\,\tilde{c}_i^\top \tilde{c}_i - \Sigma = 0$. The sequence of nonnegative numbers $\{\lambda_{\min}(A_0^\top A_0),\ m = 1, 2, \ldots\}$ never decreases and tends to $+\infty$. Then, by the Law of large numbers in [16, Theorem 6.6, page 209]

$$\frac{1}{\lambda_{\min}(A_0^\top A_0)} \sum_{i=1}^m \big(\tilde{c}_i^\top \tilde{c}_i - \Sigma\big) \to 0 \quad \text{as } m \to \infty, \quad \text{a.s.,}$$

whence, with (58),

$$\|M_2\| \leq \frac{\|\sum_{i=1}^m (\tilde{c}_i^\top \tilde{c}_i - \Sigma)\|}{\lambda_{\min}(A_0^\top A_0)} \to 0 \quad \text{as } m \to \infty, \quad \text{a.s.;}$$

$$M_2 \to 0 \quad \text{as } m \to \infty, \quad \text{a.s.}$$ □

8.4 Proof of the uniqueness theorems

Proof of Theorem 4.1. The random events 1, 2 and 3 are defined in the statement of this theorem on page 256. The random event 1 always occurs. This was proved in Section 2.2 where the estimator $\widehat{X}_{\mathrm{ext}}$ is defined. In order to prove the rest, we first construct the random event (59), which occurs either with high probability or eventually. Then we prove that, whenever (59) occurs, there is the existence and "more than uniqueness" in the random event 3, and then prove that the random event 2 occurs.

Now, we construct a modified version $\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}$ of the estimator $\widehat{X}_{\mathrm{ext}}$ in the following way. If there exist such solutions $(\Delta, \widehat{X}_{\mathrm{ext}})$ to (7) & (8) that $\|\sin\angle(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0)\| \geq (1+\|X_0\|^2)^{-1/2}$, let $\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}$ come from one of such solutions. Otherwise, if for every solution $(\Delta, \widehat{X}_{\mathrm{ext}})$ to (7) & (8) $\|\sin\angle(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0)\| < (1+\|X_0\|^2)^{-1/2}$, let $\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}$ come from one of these solutions. In any case, let us construct $\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}$ in such a way that it is a random matrix. It is possible; that follows from [17].

Thus we construct a matrix $\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}$ such that:

1. $\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}$ is a $(d+n)\times n$ random matrix;
2. for some $\Delta \in \mathbb{R}^{m\times(d+n)}$, $(\Delta, \widehat{X}_{\mathrm{ext}}^{\mathrm{mod}})$ is a solution to (7) & (8);
3. if $\|\sin\angle(\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}, X_{\mathrm{ext}}^0)\| < (1+\|X_0\|^2)^{-1/2}$, then $\|\sin\angle(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0)\| < (1+\|X_0\|^2)^{-1/2}$ for any solution $(\Delta, \widehat{X}_{\mathrm{ext}})$ to (7) & (8).

From the proof of Theorem 3.5 it follows that $\|\sin\angle(\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}, X_{\mathrm{ext}}^0)\| \to 0$ in probability as $m\to\infty$. From the proof of Theorem 3.6 or 3.7 it follows that $\|\sin\angle(\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}, X_{\mathrm{ext}}^0)\| \to 0$ almost surely. Then

$$\left\|\sin\angle\big(\widehat{X}_{\mathrm{ext}}^{\mathrm{mod}}, X_{\mathrm{ext}}^0\big)\right\| < \frac{1}{\sqrt{1+\|X_0\|^2}} \tag{59}$$

either with high probability or almost surely.

Whenever the random event (59) occurs, for any solution Δ to (7) and the corresponding full-rank solution $\widehat{X}_{\mathrm{ext}}$ to (8) (which always exists) it holds that $\|\sin\angle(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0)\| < (1+\|X_0\|^2)^{-1/2}$, whence, due to Theorem 8.3, the bottom $d\times d$ block of the matrix $\widehat{X}_{\mathrm{ext}}$ is nonsingular. Right-multiplying $\widehat{X}_{\mathrm{ext}}$ by a nonsingular matrix, we can transform it into a form $\left(\begin{smallmatrix}\widehat{X}\\ -I\end{smallmatrix}\right)$. The constructed matrix $\widehat{X}$ is a solution to equation (9) for given Δ. Thus, we have just proved that if the random event (59) occurs, then for any Δ which is a solution to (7), equation (9) has a solution.

Now, prove the uniqueness of $\widehat{X}$. Let $(\Delta_1, \widehat{X}_1)$ and $(\Delta_2, \widehat{X}_2)$ be two solutions to (7) & (9). Show that $\widehat{X}_1 = \widehat{X}_2$. (If we can for $\Delta_1 = \Delta_2$, then the random event 3 occurs.) Denote $\widehat{X}_1^{\mathrm{ext}} = \left(\begin{smallmatrix}\widehat{X}_1\\ -I\end{smallmatrix}\right)$ and $\widehat{X}_2^{\mathrm{ext}} = \left(\begin{smallmatrix}\widehat{X}_2\\ -I\end{smallmatrix}\right)$. By Proposition 7.9, $\operatorname{span}\langle\widehat{X}_1^{\mathrm{ext}}\rangle \subset \operatorname{span}\langle u_k,\ \nu_k\leq d\rangle$ and $\operatorname{span}\langle\widehat{X}_2^{\mathrm{ext}}\rangle \subset \operatorname{span}\langle u_k,\ \nu_k\leq d\rangle$, where ν_k and u_k are generalized eigenvalues (arranged in ascending order) and respective eigenvectors of the matrix pencil $\langle X^\top X,\ \Sigma\rangle$.

Assume by contradiction that $\widehat{X}_1 \neq \widehat{X}_2$. Then $\operatorname{rk}[\widehat{X}_1^{\mathrm{ext}},\ \widehat{X}_2^{\mathrm{ext}}] \geq d+1$, where $[\widehat{X}_1^{\mathrm{ext}},\ \widehat{X}_2^{\mathrm{ext}}]$ is an $(n+d)\times 2d$ matrix constructed of $\widehat{X}_1^{\mathrm{ext}}$ and $\widehat{X}_2^{\mathrm{ext}}$. Then

$$d^* = \operatorname{rk}\langle u_k,\ \nu_k\leq d\rangle \geq \operatorname{rk}\left[\widehat{X}_1^{\mathrm{ext}},\quad \widehat{X}_2^{\mathrm{ext}}\right] \geq d+1$$

(which means $\nu_d = \nu_{d+1}$). Then $d_*-1 < d < d^*$, where $d_*-1 = \dim\operatorname{span}\langle u_k,\ \nu_k < d\rangle$, $d = \dim\operatorname{span}\langle X_{\mathrm{ext}}^0\rangle$ and $d^* = \dim\operatorname{span}\langle u_k,\ \nu_k\leq d\rangle$ (notation d_* and d^* comes from the proof of Proposition 7.9). By Lemma 6.4, there exists a d-dimensional subspace V_{12} for which $\operatorname{span}\langle u_k,\ \nu_k<d\rangle \subset V_{12} \subset \operatorname{span}\langle u_k,\ \nu_k\leq d\rangle$ and $\|\sin\angle(V_{12}, X_{\mathrm{ext}}^0)\| = 1$. Bind a basis of the d-dimensional subspace $V_{12}\subset\mathbb{R}^{(n+d)}$ into the $(n+d)\times d$ matrix $\widehat{X}_3^{\mathrm{ext}}$, so $\operatorname{span}\langle\widehat{X}_3^{\mathrm{ext}}\rangle = V_{12}$. Again, by Proposition 7.9 for some matrix Δ, $(\Delta, \widehat{X}_3^{\mathrm{ext}})$ is a solution to (7) & (9). Then $\|\sin\angle(\widehat{X}_3^{\mathrm{ext}},$

$X^0_{\mathrm{ext}})\| = 1 \geq (1+\|X_0\|^2)^{-1/2}$. Then $\|\sin\angle(\widehat{X}^{\mathrm{mod}}_{\mathrm{ext}}, X^0_{\mathrm{ext}})\| \geq (1+\|X_0\|^2)^{-1/2}$, which contradicts (59). Thus, the random event 3 occurs.

Now prove that the random event 2 occurs. Let Δ_1 and Δ_2 be two solutions to the optimization problem (7). Whenever the random event (59) occurs, the respective solutions $\widehat{X}_1$ and $\widehat{X}_2$ to equation (9) exist. By already proved uniqueness, they are equal, i.e., $\widehat{X}_1 = \widehat{X}_2$. Then both Δ_1 and Δ_2 are solutions to the optimization problem

$$\begin{cases} \|\Delta\,(\Sigma^{1/2})^{\dagger}\|_F \to \min; \\ \Delta\,(I - P_\Sigma) = 0; \\ (C-\Delta)\widehat{X}_1^{\mathrm{ext}} = 0 \end{cases} \tag{60}$$

for the fixed $\widehat{X}_1^{\mathrm{ext}} = \left(\begin{smallmatrix}\widehat{X}_1 \\ -I\end{smallmatrix}\right) = \left(\begin{smallmatrix}\widehat{X}_2 \\ -I\end{smallmatrix}\right)$. By Proposition 7.2 and Remark 7.2-1, the least element in the optimization problem (28) for $X = \widehat{X}_1^{\mathrm{ext}}$ is attained for the unique matrix $\Delta = C\widehat{X}_1^{\mathrm{ext}}(\widehat{X}_1^{\mathrm{ext}\top}\Sigma\widehat{X}_1^{\mathrm{ext}})^{\dagger}\widehat{X}_1^{\mathrm{ext}\top}\Sigma$. Since it is attained, it is also attained for both Δ_1 and Δ_2. Hence, $\Delta_1 = \Delta_2$. Thus, the random event 2 occurs.

We proved that the random event 1 always occurs, and the random events 2 and 3 occur whenever (59) occurs, which occurs either with high probability or eventually as desired. □

Remark 8.1. This uniqueness of the solution Δ to the optimization problem (7) agrees with the uniqueness result in [6]. The solution is unique if $\nu_d < \nu_{d+1}$.

Proof of Theorem 4.2. 1. In Theorem 4.1, the event 1 occurs always, not just with high probability or eventually. The solution Δ to (7) exists and also solves (11) due to Proposition 7.6. Thus, the first sentence of Theorem 4.2 is true. The second sentence of Theorem 4.2 has been already proved, since the constraints in the optimization problems (7) and (11) are the same.

2 & 3. The proof of consistency of the estimator defined with (11) & (9) and of the existence of the solution is similar to the proof for the estimator defined with (7) & (9) in Theorems 3.5–3.7 and 4.1. The only difference is skipping the use of Proposition 7.6. Notice that we do not prove the uniqueness of the solution because we cannot use Proposition 7.9. □

To Remark 4.2-1. The amended Theorem 4.2 can be proved similarly. In the proof of part 1, read "The solution Δ to (7) ... solves (12) due to Proposition 7.8." In the proof of parts 2 and 3, read "The only difference is using Proposition 7.8, part 2 instead of Proposition 7.6."

Proofs of auxiliary results

8.5 *Proof of lemmas on perturbation bounds for invariant subspaces*

Proof of Lemma 6.5 and Remark 6.5-1. For the proof of Lemma 6.5 itself, see parts 2 and 3 of the proof below. For the proof of Remark 6.5-1, see parts 2, 3 and 4 below. Part 1 is a mere discussion of why the conditions of Remark 6.5-1 are more general than ones of Lemma 6.5.

In the proof, we assume that $\{x : x^\top Bx > 0\}$ is the domain of the function $f(x)$. The assumption affects the definition of $\lim_{x\to x_*} f(x)$, and $\inf f$ is the infimum of $f(x)$ *over the domain.*

1. At first, clarify the conditions of Remark 6.5-1. As it is, the existence of a point x such that

$$\liminf_{\vec{t}\to x} f(\vec{t}) = \inf_{\vec{t}^\top B\vec{t}>0} f(\vec{t}) \tag{61}$$

is assumed in Remark 6.5-1. Now, prove that, under the preceding condition of Remark 6.5-1, there exists a vector $x \neq 0$ that satisfies (61).

The function $f(x)$ is homogeneous of degree 0, i.e.,

$$f(kx) = f(x) \quad \text{if } k \in \mathbb{R} \setminus \{0\} \text{ and } x^\top Bx > 0.$$

Hence, all values which are attained by $f(x)$ on its domain $\{x : x^\top Bx > 0\}$, are also attained on the bounded set $\{x : \|x\|=1,\ x^\top Bx > 0\}$:

$$f\big(\{x : \|x\|=1,\ x^\top Bx > 0\}\big) = f\big(\{x : x^\top Bx > 0\}\big).$$

Then

$$\inf_{\substack{\|x\|=1\\ x^\top Bx>0}} f(x) = \inf_{x^\top Bx>0} f(x).$$

Let F be a closure of $\{x : \|x\|=1,\ x^\top Bx > 0\}$. There is a sequence $\{x_k,\ k = 1, 2, \ldots\}$ such that $\|x_k\|=1$ and $x_k^\top Bx_k > 0$ for all k, and $\lim_{k\to\infty} f(x_k) = \inf_{x^\top Bx>0} f(x)$. Since F is a compact set, there exists $x_* \in F$ which is a limit of some subsequence $\{x_{k_i},\ i = 1, 2, \ldots\}$ of $\{x_k,\ k = 1, 2, \ldots\}$. Then either

$$\liminf_{x\to x_*} f(x) \le \inf_{x^\top Bx>0} f(x) \tag{62}$$

or, if $x_{k_i} = x_*$ for i large enough,

$$f(x_*) \le \inf_{x^\top Bx>0} f(x). \tag{63}$$

(In equations (62) and (63), we assume that $\{x : x^\top Bx > 0\}$ is a domain of $f(x)$, so (63) implies $x_*^\top Bx_* > 0$.) Again, due to the homogeneity, $\liminf\limits_{x\to x_*} f(x) \le f(x_*)$ if $f(x_*)$ makes sense. Hence (62) follows from (63) and thus holds true either way.

Taking the limit in the relation $f(x) \ge \inf f$, we obtain the opposite inequality

$$\liminf_{x\to x_*} f(x) \ge \inf_{x^\top Bx>0} f(x).$$

Thus, the equality (25) holds true for some $x_* \in F$. Note that $\|x_*\| = 1$, so $x_* \neq 0$.

2. Prove that under the conditions of Lemma 6.5 or Remark 6.5-1

$$\left[\begin{array}{ll} \text{either} & f(x_*) \le f(x) \\ \text{or} & x_*^\top (A + \tilde{A})x_* \le 0. \end{array}\right.$$

Because the matrix B is symmetric and positive semidefinite, $x^\top Bx = 0$ if and only if $Bx = 0$, and $x^\top Bx > 0$ if and only if $Bx \neq 0$. As $Bx_0 \neq 0$, $x_0^\top Bx_0 > 0$ and the function $f(x)$ is well-defined at x_0.

Under the conditions of Lemma 6.5 the function $f(x)$ is well-defined at x_0 and attains its minimum at x_*, so $f(x_*) \le f(x_0)$.

Under the conditions of Remark 6.5-1 we consider 3 cases concerning the value of $x_*^\top Bx_*$.

Case 1. $x_*^\top B x_* < 0$. But on the domain of $f(x)$ the inequality $x^\top B x > 0$ holds true. Since x_* is a limit point of the domain of $f(x)$, the inequality $x_*^\top B x_* \ge 0$ holds true, and Case 1 is impossible.

Case 2. $x_*^\top B x_* = 0$. Prove that $x_*^\top (A + \tilde{A}) x_* \le 0$. On the contrary, let $x_*^\top (A + \tilde{A}) x_* > 0$. Remember once again that $x^\top B x > 0$ on the domain of $f(x)$. Then

$$\lim_{x \to x_*} f(x) = \lim_{x \to x_*} \frac{x^\top (A + \tilde{A}) x}{x^\top B x} = +\infty,$$

which cannot be $\inf f(x)$. The contradiction obtained implies that $x_*^\top (A + \tilde{A}) x_* \le 0$.

Case 3. $x_*^\top B x_* > 0$. Then the function $f(x)$ is well-defined at x_*, and

$$f(x_*) = \lim_{x \to x_*} f(x) = \inf f(x) \le f(x_0).$$

So, $f(x_*) \le f(x_0)$ in Case 3.

3. Proof of Lemma 6.5 and proof of Remark 6.5-1 when $f(x_*) \le f(x_*)$. Then

$$\frac{x^\top (A + \tilde{A}) x}{x^\top B x} \le \frac{x_0^\top (A + \tilde{A}) x_0}{x_0^\top B x_0} .$$

As $A x_0 = 0$,

$$x^\top A x \le -x^\top \tilde{A} x + \frac{x_0^\top \tilde{A} x_0 \, x^\top B x}{x_0^\top B x_0} \le \|\tilde{A}\| \left(\|x\|^2 + \frac{\|x_0\|^2 x^\top B x}{x_0^\top B x_0} \right).$$

With use of eigendecomposition of A, the inequality $x^\top A x \ge \lambda_2(A) \, \|x\|^2 \times \sin^2 \angle(x, x_0)$ can be proved. Hence the desired inequality follows:

$$\lambda_2(A) \sin^2 \angle(x, x_0) \le \|\tilde{A}\| \left(1 + \frac{\|x_0\|^2}{x_0^\top B x_0} \cdot \frac{x^\top B x}{\|x\|^2} \right).$$

4. Proof of Remark 6.5-1 when $x_*^\top (A + \tilde{A}) x_* \le 0$. Then

$$\begin{aligned} x^\top A x &\le -x^\top \tilde{A} x, \\ \lambda_2(A) \|x\|^2 \sin^2 \angle(x, x_0) &\le \|\tilde{A}\| \, \|x\|^2, \\ \lambda_2(A) \sin^2 \angle(x, x_0) &\le \|\tilde{A}\|, \end{aligned}$$

whence the desired inequality follows. □

Notation. If A and B are symmetric matrices of the same size, and furthermore the matrix B is positive definite, denote

$$\max \frac{A}{B} = \lambda_{\max}\big(B^{-1} A\big).$$

The notation is used in the proof of Lemma 6.6.

Lemma 8.2. *Let $1 \le d_1 \le n$, $0 \le d_2 \le n$. Let $X \in \mathbb{R}^{n\times d_1}$ be a matrix of full rank, and V be a d_2-dimensional subspace in $\mathbb{R}^n$. Then*

$$\max \frac{X^\top (I - P_V)X}{X^\top X} = \big\| \sin\angle(X, V)\big\|^2 \quad \textit{if} \quad d_1 \le d_2,$$
$$\max \frac{X^\top (I - P_V)X}{X^\top X} = 1 \quad \textit{if} \quad d_1 > d_2.$$

Proof. Using the min-max theorem, the relation $\operatorname{span}\langle X\rangle = \operatorname{span}\langle P_{\operatorname{span}\langle X\rangle}\rangle$ and simple properties of orthogonal projectors, construct the inequality

$$\begin{aligned}
&\max \frac{X^\top (I - P_V)X}{X^\top X}\\
&\quad= \max_{v\in\mathbb{R}^{d_1}\setminus\{0\}} \frac{v^\top X^\top (I-P_V)Xv}{v^\top X^\top X v}\\
&\quad= \max_{w\in\operatorname{span}\langle X\rangle\setminus\{0\}} \frac{w^\top (I-P_V)w}{w^\top w} = \max_{v\in\mathbb{R}^n\setminus\{0\}} \frac{v^\top P_{\operatorname{span}\langle X\rangle}(I-P_V)P_{\operatorname{span}\langle X\rangle}v}{v^\top P_{\operatorname{span}\langle X\rangle}P_{\operatorname{span}\langle X\rangle}v}\\
&\quad\ge \max_{v\in\mathbb{R}^n\setminus\{0\}} \frac{v^\top P_{\operatorname{span}\langle X\rangle}(I-P_V)P_{\operatorname{span}\langle X\rangle}v}{v^\top v} = \lambda_{\max}\big(P_{\operatorname{span}\langle X\rangle}(I-P_V)P_{\operatorname{span}\langle X\rangle}\big)\\
&\quad= \lambda_{\max}\big(P_{\operatorname{span}\langle X\rangle}(I-P_V)(I-P_V)P_{\operatorname{span}\langle X\rangle}\big) = \big\|P_{\operatorname{span}\langle X\rangle}(I-P_V)\big\|^2.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\max_{w\in\operatorname{span}\langle X\rangle\setminus\{0\}} \frac{w^\top (I-P_V)w}{w^\top w} &= \max_{w\in\operatorname{span}\langle X\rangle\setminus\{0\}} \frac{w^\top P_{\operatorname{span}\langle X\rangle}(I-P_V)P_{\operatorname{span}\langle X\rangle}w}{w^\top w}\\
&\le \max_{v\in\mathbb{R}^n\setminus\{0\}} \frac{v^\top P_{\operatorname{span}\langle X\rangle}(I-P_V)P_{\operatorname{span}\langle X\rangle}v}{v^\top v}.
\end{aligned}$$

Thus,

$$\max \frac{X^\top (I - P_V)X}{X^\top X} = \big\|P_{\operatorname{span}\langle X\rangle}(I-P_V)\big\|^2.$$

If $d_1 \le d_2$, then $\|P_{\operatorname{span}\langle X\rangle}(I-P_V)\| = \|\sin\angle(X,V)\|$ due to (23). Otherwise, if $d_1 > d_2$, then

$$\dim\operatorname{span}\langle X\rangle + \dim V^\perp = \operatorname{rk} X + n - \dim V = d_1 + n - d_2 > n.$$

Hence the subspaces $\operatorname{span}\langle X\rangle$ and $V^\perp$ have nontrivial intersection, i.e., there exists $w \ne 0$, $w \in \operatorname{span}\langle X\rangle \cap V^\perp$. Then $P_{\operatorname{span}\langle X\rangle}(I-P_V)w = w$, whence $\|P_{\operatorname{span}\langle X\rangle}(I-P_V)\| \ge 1$. On the other hand, $\|P_{\operatorname{span}\langle X\rangle}(I-P_V)\| \le \|P_{\operatorname{span}\langle X\rangle}\| \times \|(I-P_V)\| \le 1$. Thus, $\|P_{\operatorname{span}\langle X\rangle}(I-P_V)\| = 1$. This completes the proof. □

Proof of Lemma 6.6. The matrix B is positive semidefinite, the matrix $X_0^\top B X_0$ is positive definite, and the matrix X_0 is of full rank d (hence, $n \ge d$). The matrix A satisfies inequality $A \ge \lambda_{d+1}(A)(I - P_{\operatorname{span}\langle X_0\rangle})$ in the Loewner order.

Let X be a point where the functional $f(x)$ defined in (26) attains its minimum. Since $X_0^\top B X_0$ is positive definite, $f(X_0)$ makes sense. Thus, $f(X) \le f(X_0)$,

$$\max \frac{X^\top (A+\tilde{A})X}{X^\top B X} \le \max \frac{X_0^\top (A+\tilde{A})X_0}{X_0^\top B X_0}.$$

Using the relations

$$\begin{aligned} X^\top \tilde{A} X &\ge -\|\tilde{A}\|\, X^\top X, & X_0^\top \tilde{A} X_0 &\le \|\tilde{A}\|\, X_0^\top X_0, \\ X^\top B X &\le \|B\|\, X^\top X, & A X_0 &= 0, \end{aligned}$$

we have

$$\max \frac{X^\top A X - \|\tilde{A}\| X^\top X}{\|B\|\, X^\top X} \le \max \frac{\|\tilde{A}\|\, X_0^\top X_0}{X_0^\top B X_0},$$

$$\frac{1}{\|B\|} \cdot \left(\max \frac{X^\top A X}{X^\top X} - \|\tilde{A}\| \right) \le \|\tilde{A}\| \max \frac{X_0^\top X_0}{X_0^\top B X_0}. \tag{64}$$

Since $A \ge \lambda_{d+1}(A)(I - P_{\mathrm{span}\langle X_0 \rangle})$, by Lemma 8.2

$$\lambda_{d+1}(A) \left\| \sin \angle(X, X_0) \right\|^2 \le \lambda_{d+1}(A) \max \frac{X^\top (I - P_{\mathrm{span}\langle X_0 \rangle})}{X^\top X} \le \max \frac{X^\top A X}{X^\top X}.$$

Then the desired inequality follows from (64):

$$\left\| \sin \angle(X, X_0) \right\|^2 \le \frac{\|\tilde{A}\|}{\lambda_{d+1}(A)} \left(1 + \|B\| \max \frac{X_0^\top X_0}{X_0^\top B X_0} \right). \qquad \square$$

8.6 *Comparison of* $\| \sin \angle(\widehat{X}_{\mathrm{ext}}, X_{\mathrm{ext}}^0) \|$ *and* $\|\widehat{X} - X_0\|$

In the next theorem and in its proof, matrices A, B and Σ have different meaning than elsewhere in the paper.

Theorem 8.3. *Let* $\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right)$ *and* $\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)$ *be full-rank* $(n+d) \times d$ *matrices. If*

$$\left\| \sin \angle \left(\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} X_0 \\ -I \end{pmatrix} \right) \right\| < \frac{1}{\sqrt{1 + \|X_0\|^2}}, \tag{65}$$

then:

1) the matrix B *is nonsingular;*

2) $\|AB^{-1} + X_0\| \le \frac{(1+\|X_0\|^2)\,(\|X_0\| s^2 + s\sqrt{1-s^2})}{1-(1+\|X_0\|^2)\, s^2}$ *with* $s = \| \sin \angle(\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right), \left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right))\|$.

Proof. *1.* Split the matrix $P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)}$, which is an orthogonal projector along the column space of the matrix $\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)$, into four blocks:

$$I - P_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} = P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} = \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \\ \mathbf{P}_2^\top & \mathbf{P}_4 \end{pmatrix}.$$

Up to the end of the proof, $\mathbf{P}_1$ means the upper-left $n \times n$ block of the $(n+p) \times (n+p)$ matrix $P^{\perp}_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)}$. Prove that $\lambda_{\min}(\mathbf{P}_1) = \frac{1}{1+\|X_0\|^2}$.

Let $X_0 = U \Sigma V^\top$ be a singular value decomposition of the matrix X_0 (here Σ is a diagonal $n \times d$ matrix, U and V are orthogonal matrices). Then

$$\begin{aligned} P^{\perp}_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} &= I - \begin{pmatrix} X_0 \\ -I \end{pmatrix} \left(\begin{pmatrix} X_0 \\ -I \end{pmatrix}^\top \begin{pmatrix} X_0 \\ -I \end{pmatrix} \right)^{-1} \begin{pmatrix} X_0 \\ -I \end{pmatrix} \\ &= \begin{pmatrix} U(I - \Sigma(\Sigma^\top \Sigma + 1)^{-1} \Sigma^\top) U^\top & U \Sigma (\Sigma^\top \Sigma + I)^{-1} V^\top \\ V(\Sigma^\top \Sigma + I)^{-1} \Sigma^\top U^\top & V(I - (\Sigma^\top \Sigma + I)^{-1}) V^\top \end{pmatrix}. \end{aligned}$$

The $n \times n$ matrix $I - \Sigma(\Sigma^\top \Sigma + I)^{-1} \Sigma^\top$ is diagonal; its diagonal entries are $\frac{1}{1+\sigma_i^2(X_0)}$, $i = 1, \ldots, n$, where

$\sigma_i(X_0)$ is the i-th singular value of X_0 if $1 \le i \le \min(n, d)$,
$\sigma_i(X_0) = 0$ if $\min(n, d) < i \le n$.

Those diagonal entries comprise all the eigenvalues of $\mathbf{P}_1$;

$$\lambda_{\min}(\mathbf{P}_1) = \frac{1}{1 + \sigma^2_{\max}(\|X_0\|)} = \frac{1}{1 + \|X_0\|^2}.$$

2. Due to equation (23), the square of the largest of sines of canonical eigenvalues between the subspaces V_1 and V_2 is equal to

$$\left\| \sin \angle(V_1, V_2) \right\|^2 = \max_{v \in V_1 \setminus \{0\}} \frac{v^\top P^{\perp}_{V_2} v}{\|v\|^2}.$$

Hence for $v \in V_1$, $v \ne 0$,

$$\left\| \sin \angle(V_1, V_2) \right\|^2 \ge \frac{v^\top P^{\perp}_{V_2} v}{\|v\|^2}. \tag{66}$$

3. Prove the first statement of Theorem 8.3 by contradiction. Suppose that the matrix B is singular. Then there exist $f \in \mathbb{R}^d \setminus \{0\}$ and $u = Af \in \mathbb{R}^n$ such that $Bf = 0$ and

$$\begin{pmatrix} u \\ 0_{d \times 1} \end{pmatrix} = \begin{pmatrix} Af \\ Bf \end{pmatrix} \in V_1,$$

where $V_1 \subset \mathbb{R}^{n+d}$ is the column space of the matrix $\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right)$. As the columns of the matrix $\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right)$ are linearly independent, $\left(\begin{smallmatrix} u \\ 0 \end{smallmatrix}\right) \ne 0$. Then, by (66),

$$\begin{aligned} \left\| \sin \angle \left(\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} X_0 \\ -I \end{pmatrix} \right) \right\|^2 \ge \frac{\begin{pmatrix} u \\ 0 \end{pmatrix}^\top P^{\perp}_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} \begin{pmatrix} u \\ 0 \end{pmatrix}}{\|\left(\begin{smallmatrix} u \\ 0 \end{smallmatrix}\right)\|^2} = \frac{u^\top \mathbf{P}_1 u}{\|u\|^2} \ge \\ \ge \lambda_{\min}(\mathbf{P}_1) = \frac{1}{1 + \|X_0\|^2}, \end{aligned}$$

which contradicts condition (65).

4. Prove inequality (67). (Later on we will show that the second statement of Theorem 8.3 follows from (67)). There exists such a vector $f \in \mathbb{R}^d \setminus \{0\}$ that $\|(AB^{-1} + X_0) f\| = \|AB^{-1} + X_0\| \, \|f\|$. Denote

$$u = \big(AB^{-1} + X_0\big) f,$$
$$z = \begin{pmatrix} A \\ B \end{pmatrix} B^{-1} f = \begin{pmatrix} AB^{-1} f \\ f \end{pmatrix} = \begin{pmatrix} u \\ 0 \end{pmatrix} - \begin{pmatrix} X_0 \\ -I \end{pmatrix} f \in V_1.$$

Since $(X_0^\top, -I) P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} = 0$ and $P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} \left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right) = 0$,

$$\begin{aligned} z^\top P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} z &= \left(\begin{pmatrix} u \\ 0 \end{pmatrix} - \begin{pmatrix} X_0 \\ -I \end{pmatrix} f\right)^\top P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} \left(\begin{pmatrix} u \\ 0 \end{pmatrix} - \begin{pmatrix} X_0 \\ -I \end{pmatrix} f\right) \\ &= \begin{pmatrix} u \\ 0 \end{pmatrix}^\top P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} \begin{pmatrix} u \\ 0 \end{pmatrix} = u^\top \mathbf{P}_1 u \\ &\geq \|u\|^2 \lambda_{\min}(\mathbf{P}_1) = \frac{\|AB^{-1} + X_0\|^2 \, \|f\|^2}{1 + \|X_0\|^2}. \end{aligned}$$

Notice that $z \neq 0$ because $B^{-1} f \neq 0$ and the columns of the matrix $\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right)$ are linearly independent. Thus,

$$0 < \|z\|^2 = \left\|AB^{-1} f\right\|^2 + \left\|f^2\right\| \leq \left(1 + \left\|AB^{-1}\right\|^2\right) \|f\|^2.$$

By (66),

$$\left\|\sin \angle\left(\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} X_0 \\ -I \end{pmatrix}\right)\right\|^2 \geq \frac{z^\top P^\perp_{\left(\begin{smallmatrix} X_0 \\ -I \end{smallmatrix}\right)} z}{\|z\|^2} \geq \frac{\|AB^{-1} + X_0\|^2}{(1 + \|X_0\|^2)\,(1 + \|AB^{-1}\|^2)},$$
$$\left\|\sin \angle\left(\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} X_0 \\ -I \end{pmatrix}\right)\right\| \geq \frac{\|AB^{-1} + X_0\|}{\sqrt{1 + \|X_0\|^2}\,\sqrt{1 + (\|X_0\| + \|AB^{-1} + X_0\|)^2}}. \tag{67}$$

5. Prove that the second statement of Theorem 8.3 follows from (67). The function

$$s(\delta) := \frac{\delta}{\sqrt{1 + \|X_0\|^2}\,\sqrt{1 + (\|X_0\| + \delta)^2}} \tag{68}$$

is strictly increasing on $[0, +\infty)$, with $s(0) = 0$ and $\lim_{\delta \to +\infty} s(\delta) = \frac{1}{\sqrt{1 + \|X_0\|^2}}$. Therefore, inequality (67) implies the implication:

$$\text{if } \left\|AB^{-1} + X_0\right\| > \delta,$$
$$\text{then } \left\|\sin \angle\left(\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} X_0 \\ -I \end{pmatrix}\right)\right\| > \frac{\delta}{\sqrt{1 + \|X_0\|^2}\,\sqrt{1 + (\|X_0\| + \delta)^2}}.$$

The equivalent contrapositive implication is as follows:

$$\text{if } \left\|\sin \angle\left(\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} X_0 \\ -I \end{pmatrix}\right)\right\| \leq \frac{\delta}{\sqrt{1 + \|X_0\|^2}\,\sqrt{1 + (\|X_0\| + \delta)^2}},$$

$$\text{then } \left\| AB^{-1} + X_0 \right\| \le \delta. \tag{69}$$

The inverse function to $s(\delta)$ in (68) is

$$\delta(s) := \frac{(1 + \|X_0\|^2)\,(s^2\,\|X_0\| + s\sqrt{1-s^2})}{1 - (1 + \|X_0\|^2)s^2}.$$

Substitute $\delta = \delta(\|\sin\angle((\begin{smallmatrix}A\\B\end{smallmatrix}),\,(\begin{smallmatrix}X_0\\-I\end{smallmatrix}))\|)$ into (69) and obtain the following statement:

$$\text{if } \left\|\sin\angle\left(\begin{pmatrix}A\\B\end{pmatrix}\,\begin{pmatrix}X_0\\-I\end{pmatrix}\right)\right\| \le \left\|\sin\angle\left(\begin{pmatrix}A\\B\end{pmatrix},\begin{pmatrix}X_0\\-I\end{pmatrix}\right)\right\|,$$
$$\text{then } \left\|AB^{-1} + X_0\right\| \le \delta\big(\left\|\sin\angle\big((\begin{smallmatrix}A\\B\end{smallmatrix}),(\begin{smallmatrix}X_0\\-I\end{smallmatrix})\big)\right\|\big),$$

whence the second statement of Theorem 8.3 follows.

In part 5 of the proof, condition (65) is used twice. First, it is one of conditions of the first statement of the theorem: without it, the matrix B might be singular. Second, the function $\delta(s)$ is defined on interval $[0, \frac{1}{\sqrt{1+\|X_0\|^2}})$. □

Corollary. *Let* $(\begin{smallmatrix}X_0\\-I\end{smallmatrix})$ *be an* $(n+d)\times d$ *matrix, and let* $\{(\begin{smallmatrix}A_m\\B_m\end{smallmatrix}),\quad m = 1, 2, \ldots\}$ *be a sequence of* $(n+d)\times d$ *matrices of rank* d. *If* $\|\sin\angle((\begin{smallmatrix}A_m\\B_m\end{smallmatrix}),\ (\begin{smallmatrix}X_0\\-I\end{smallmatrix}))\| \to 0$ *as* $m\to\infty$, *then:*

1) the matrix B_m *is nonsingular for* m *large enough,*

2) $-A_m B_m^{-1} \to X_0$ *as* $m\to\infty$.

8.7 Generalized eigenvalue problem for positive semidefinite matrices: proofs

Proof of Lemma 7.1. For fixed i, split the matrix T in two blocks. Let $T = [T_{i1}, T_{i2}]$, where T_{i1} is the matrix constructed of the first i columns of T, and T_{i2} is the matrix constructed of the last $n-i+1$ columns of T. Denote V_1 and V_2 the column spaces of the matrices T_{i1} and T_{i2}, respectively. Then $\dim V_1 = i$ and $\dim V_2 = n-i+1$.

1. The proof of the fact that $\nu_i \in \{\lambda \ge 0 \mid \text{“}\exists V,\ \dim V = i : (A-\lambda B)|_V \le 0\text{”}\}$ *if* $\nu_i < \infty$. In other words, if $\nu_i < \infty$, then relations

$$\lambda \ge 0, \qquad \dim(V) = i, \qquad (A-\lambda B)|_V \le 0 \tag{70}$$

hold true for $\lambda = \nu_i$ and $V = V_1$.

If $v \in V_1$, then $v = T_{i1}x$ for some $x \in \mathbb{R}^i$. Hence

$$v^\top(A - \nu_i B)v = x^\top T_{i1}^\top (A - \nu_i B) T_{i1} x$$
$$= x^\top \operatorname{diag}(\lambda_1 - \nu_i\mu_1,\ \ldots,\ \lambda_i - \nu_i\mu_1)x = \sum_{j=1}^{i} x_j^2(\lambda_j - \nu_i\mu_j).$$

The inequality $\lambda_j - \nu_i\mu_j \le 0$ holds true for all j such that either $\lambda_j = \mu_j = 0$ or $\lambda_j/\mu_j \le \nu_i$; particularly, it holds true for $j = 1, \ldots, i$. Hence $v^\top(A - \nu_i B)v \le 0$.

2. The proof of the fact that ν_i is a lower bound of the set $\{\lambda \geq 0 \mid$ "$\exists V,\ \dim V = i : (A - \lambda B)|_V \leq 0$"$\}$. In other words, if there exists a subspace $V \subset \mathbb{R}^n$ such that the relations (70) hold true, then $\nu_i \leq \lambda$.

By contradiction, suppose that $\dim V = i$, $(A - \lambda B)|_V \leq 0$, $\nu_i > \lambda \geq 0$. Then $\nu_i > 0$.

Now prove that $(A - \lambda B)|_{V_2} > 0$. If $v \in V_2 \setminus \{0\}$, then $v = T_{i2}x$ for some $x \in \mathbb{R}^{n-i+1} \setminus \{0\}$. Then

$$v^\top (A - \lambda B)v = \sum_{j=i}^{n} x_{j+1-i}^2 (\lambda_j - \lambda \mu_j).$$

For $j \geq i$, due to the inequality $\nu_j \geq \nu_i > 0$ and the conditions of the lemma, the case $\lambda_j = 0$ is impossible; thus $\lambda_j > 0$. Prove the inequality $\lambda_j - \lambda\mu_j > 0$. If $\mu_j > 0$, then $\lambda_j - \lambda\mu_j = (\nu_j - \lambda)\mu_j$. Since $\nu_j \geq \nu_i > \lambda$, the first factor $\nu_i - \lambda$ is a positive number. Hence, $\lambda_j - \lambda\mu_j > 0$. Otherwise, if $\mu_j = 0$, then $\lambda_j - \lambda\mu_j = \lambda_j > 0$. Thus the inequality $\lambda_j - \lambda\mu_j > 0$ holds true in both cases. Hence $v^\top (A - \lambda B)v > 0$. Since this holds for all $v \in V_2 \setminus \{0\}$, the restriction of the quadratic form $A - \lambda B$ onto the linear subspace V_2 is positive definite.

On the one hand, since $(A - \lambda B)|_V \leq 0$ and $(A - \lambda B)|_{V_2} > 0$, the subspaces V and V_2 have a trivial intersection. On the other hand, since $\dim V + \dim V_2 = n + 1 > n$, the subspaces V and V_2 cannot have a trivial intersection. We got a contradiction.

Hence $\nu_i \leq \lambda$, and ν_i is a lower bound of $\{\lambda \geq 0 \mid$ "$\exists V,\ \dim V = i : (A - \lambda B)|_V \leq 0$"$\}$. That completes the proof of Lemma 7.1. □

Remember that $M^\dagger$ is the Moore–Penrose pseudoinverse matrix to M; $\operatorname{span}\langle M\rangle$ is the column span of the matrix M. If matrices M and N are compatible for multiplication, then $\operatorname{span}\langle MN\rangle \subset \operatorname{span}\langle M\rangle$. (Furthermore, $\operatorname{span}\langle M_1\rangle \subset \operatorname{span}\langle M_2\rangle$ if and only if $M_1 = M_2 N$ for some matrix N). Hence, $\operatorname{span}\langle MM^\top\rangle = \operatorname{span}\langle M\rangle$ (to prove it, we can use the identity $M = MM^\top (M^\top)^\dagger$).

Since the $n \times n$ covariance matrix Σ is positive semidefinite, for every $k \times n$ matrix M the equality $\operatorname{span}\langle M\Sigma M^\top\rangle = \operatorname{span}\langle M\Sigma\rangle$ holds true. This can be proved with use of the matrix square root.

If what follows, for a fixed $(n + d) \times d$ matrix X denote

$$\Delta_{\mathrm{pm}} = CX\big(X^\top \Sigma X\big)^\dagger X^\top \Sigma,$$

where C is an $m \times (n + d)$ matrix, Σ is an $n \times n$ positive semidefinite matrix.

Proof of Proposition 7.2. *1, necessity. Relation* (30) *is a necessary condition for compatibility of the constraints in* (28). Let $\Delta\,(I - P_\Sigma) = 0$ and $(C - \Delta)X = 0$ for some $m \times (n + d)$ matrix Δ. Due to $\Delta\,(I - P_\Sigma) = 0$, $\Delta = M\Sigma$ for some matrix M. Then $CX = \Delta X = M\Sigma X$, $X^\top C^\top = X^\top \Sigma M^\top$, whence $\operatorname{span}(X^\top C^\top) \subset \operatorname{span}(X^\top \Sigma)$.

1, sufficiency. Relation (30) *is a sufficient condition for compatibility of the constraints in* (28). Let $\operatorname{span}(X^\top C^\top) \subset \operatorname{span}(X^\top \Sigma)$. Then $X^\top C^\top = X^\top \Sigma M$ for some matrix M. The constraints $\Delta\,(I - P_\Sigma) = 0$, $(C - \Delta)X = 0$ are satisfied for $\Delta = M^\top \Sigma$, so they are compatible.

2a, eqns. (31)*. If the constraints are compatible, they are satisfied for* $\Delta = \Delta_{\mathrm{pm}}$. Indeed,

$$\Delta_{\mathrm{pm}}\,(I - P_\Sigma) = CX\big(X^\top \Sigma X\big)^\dagger X^\top \Sigma\,(I - P_\Sigma) = 0,$$

since $\Sigma\,(I - P_\Sigma) = 0$. If the constraints are compatible, then

$$\operatorname{span}\big(X^\top \Sigma X\big) = \operatorname{span}\big(X^\top \Sigma\big) \subset \operatorname{span}\big(X^\top C^\top\big),$$

whence

$$X^\top \Sigma X\big(X^\top \Sigma X\big)^\dagger X^\top C^\top = P_{X^\top \Sigma X} X^\top C^\top = X^\top C^\top,$$
$$\Delta_{\mathrm{pm}} X = CX\big(X^\top \Sigma X\big)^\dagger X^\top \Sigma X = CX,$$
$$(C - \Delta_{\mathrm{pm}})X = 0.$$

2a, eqn. (32) *and 2b. If the constraints are compatible, then the constrained least element of* $\Delta\Sigma^\dagger\Delta^\top$ *is attained for* $\Delta = \Delta_{\mathrm{pm}}$*. The least element is equal to* $CX(X^\top\Sigma X)^\dagger X^\top C^\top$. Let Δ satisfy the constraints, which imply $\Delta P_\Sigma = \Delta$ and $\Delta X = CX$. Expand the product

$$(\Delta - \Delta_{\mathrm{pm}})\Sigma^\dagger(\Delta - \Delta_{\mathrm{pm}})^\top = \Delta\Sigma^\dagger\Delta^\top - \Delta_{\mathrm{pm}}\Sigma^\dagger\Delta^\top - \Delta\Sigma^\dagger\Delta_{\mathrm{pm}}^\top + \Delta_{\mathrm{pm}}\Sigma^\dagger\Delta_{\mathrm{pm}}^\top. \tag{71}$$

Simplify the expressions for three (of four) summands:

$$\begin{aligned}\Delta\Sigma^\dagger\Delta_{\mathrm{pm}}^\top &= \Delta\Sigma^\dagger\Sigma X\big(X^\top\Sigma X\big)^\dagger X^\top C^\top \\ &= \Delta P_\Sigma X\big(X^\top\Sigma X\big)^\dagger X^\top C^\top \\ &= \Delta X\big(X^\top\Sigma X\big)^\dagger X^\top C^\top = CX\big(X^\top\Sigma X\big)^\dagger X^\top C^\top.\end{aligned}$$

Applying matrix transposition to both sides of the last chain of equalities, we get

$$\Delta_{\mathrm{pm}}\Sigma^\dagger\Delta^\top = CX\big(X^\top\Sigma X\big)^\dagger X^\top C^\top.$$

For the last summand,

$$\begin{aligned}\Delta_{\mathrm{pm}}\Sigma^\dagger\Delta_{\mathrm{pm}}^\top &= CX\big(X^\top\Sigma X\big)^\dagger X^\top\Sigma\Sigma^\dagger\Sigma X\big(X^\top\Sigma X\big)^\dagger X^\top C^\top \\ &= CX\big(X^\top\Sigma X\big)^\dagger X^\top\Sigma X\big(X^\top\Sigma X\big)^\dagger X^\top C^\top \\ &= CX\big(X^\top\Sigma X\big)^\dagger X^\top C^\top.\end{aligned}$$

Thus, (71) implies that

$$\Delta\Sigma^\dagger\Delta^\top = (\Delta - \Delta_{\mathrm{pm}})\Sigma^\dagger(\Delta - \Delta_{\mathrm{pm}})^\top + CX\big(X^\top\Sigma X\big)^\dagger X^\top C^\top. \tag{72}$$

Hence

$$\Delta\Sigma^\dagger\Delta^\top \geq CX\big(X^\top\Sigma X\big)^\dagger X^\top C^\top,$$

and statement 2b of the theorem is proved. For $\Delta = \Delta_{\mathrm{pm}}$, equality is attained, which coincides with (32).

Remark 7.2-1. The least point is attained for a unique Δ. It is enough to show that if Δ satisfies the constraints and $\Delta\Sigma^{\dagger}\Delta^{\top} = CX(X^{\top}\Sigma X)^{\dagger}X^{\top}C^{\top}$, then $\Delta = \Delta_{\mathrm{pm}}$.

Indeed, if Δ satisfies the constraints $\Delta\,(I - P_{\Sigma}) = 0$ and $(C-\Delta)X = 0$, and $\Delta\Sigma^{\dagger}\Delta^{\top} = CX(X^{\top}\Sigma X)^{\dagger}X^{\top}C^{\top}$, then due to (72)

$$(\Delta - \Delta_{\mathrm{pm}})\Sigma^{\dagger}(\Delta - \Delta_{\mathrm{pm}})^{\top} = 0.$$

As $\Sigma^{\dagger}$ is a positive semidefinite matrix, $(\Delta - \Delta_{\mathrm{pm}})\Sigma^{\dagger} = 0$ and $(\Delta - \Delta_{\mathrm{pm}})P_{\Sigma} = (\Delta - \Delta_{\mathrm{pm}})\Sigma^{\dagger}\Sigma = 0$. Add the equality $\Delta\,(I - P_{\Sigma}) = 0$ (which is one of the constraints) and subtract the equality $\Delta_{\mathrm{pm}}\,(I - P_{\Sigma}) = 0$ (which is one of equalities (31) and holds true due part 2a of the theorem). Obtain

$$\Delta - \Delta_{\mathrm{pm}} = (\Delta - \Delta_{\mathrm{pm}})P_{\Sigma} + \Delta\,(I - P_{\Sigma}) - \Delta_{\mathrm{pm}}\,(I - P_{\Sigma}) = 0,$$

whence $\Delta = \Delta_{\mathrm{pm}}$. □

Proof of Proposition 7.3. *1. Necessity.* Since the matrices $C^{\top}C$ and Σ are positive semidefinite, the matrix pencil $\langle C^{\top}C, \Sigma\rangle$ is definite if and only if the matrix $C^{\top}C + \Sigma$ is positive semidefinite. Thus, if the matrix pencil $\langle C^{\top}C, \Sigma\rangle$ is definite, then the matrix $C^{\top}C + \Sigma$ is positive definite. As the columns of the matrix X are linearly independent, the matrix $X(C^{\top}C + \Sigma)X^{\top} = X^{\top}C^{\top}CX + X^{\top}\Sigma X$ is positive definite as well, whence $\operatorname{span}(X^{\top}C^{\top}CX + X^{\top}\Sigma X) = \mathbb{R}^n$.

If the constraints are compatible, then the condition (30) holds true, whence

$$\begin{aligned}\mathbb{R}^n &= \operatorname{span}\langle X^{\top}C^{\top}CX + X^{\top}\Sigma X\rangle \\ &\subset \operatorname{span}\langle X^{\top}C^{\top}CX\rangle + \operatorname{span}\langle X^{\top}\Sigma X\rangle \\ &= \operatorname{span}\langle X^{\top}C^{\top}\rangle + \operatorname{span}\langle X^{\top}\Sigma\rangle \\ &= \operatorname{span}\langle X^{\top}\Sigma\rangle = \operatorname{span}\langle X^{\top}\Sigma X\rangle.\end{aligned}$$

Since $\operatorname{span}\langle X^{\top}\Sigma X\rangle = \mathbb{R}^n$, the matrix $X^{\top}\Sigma X$ is nonsingular.

2. Sufficiency. If the matrix $X^{\top}\Sigma X$ is nonsingular, then

$$\operatorname{span}\langle X^{\top}\Sigma\rangle = \operatorname{span}\langle X^{\top}\Sigma X\rangle = \mathbb{R}^n \supset \operatorname{span}\langle X^{\top}C^{\top}\rangle.$$

Thus the condition (30), which is the necessary and sufficient condition for compatibility of the constraints, holds true. □

Proof of Proposition 7.4. Construct simultaneous diagonalization of matrices $XCC^{\top}X^{\top}$ and $X\Sigma X^{\top}$ (according to Theorem 6.2) that satisfies Remark 6.2-2:

$$X^{\top}C^{\top}CX = \left(T^{-1}\right)^{\top}\Lambda T^{-1}, \qquad X^{\top}\Sigma X = \left(T^{-1}\right)^{\top}\mathrm{M}T^{-1}.$$

Notations Λ, M, $T = \begin{bmatrix}T_1 & T_2\end{bmatrix}$, μ_i, λ_i, ν_i are taken from Theorem 6.2, Remark 7.2-1, and Lemma 7.1.

The subspace

$$\operatorname{span}\langle X^{\top}C^{\top}\rangle = \operatorname{span}\langle X^{\top}C^{\top}CX\rangle = \operatorname{span}\langle\left(T^{-1}\right)^{\top}\Lambda T^{-1}\rangle = \operatorname{span}\langle\left(T^{-1}\right)^{\top}\Lambda\rangle$$

is spanned by columns of the matrix $(T^{-1})^\top$ that correspond to nonzero λ_i's. Similarly, the subspace $\operatorname{span}\langle X^\top \Sigma\rangle = \operatorname{span}\langle (T^{-1})^\top \mathrm{M}\rangle$ is spanned by columns of the matrix $(T^{-1})^\top$ that correspond to non-zero μ_i's. Note that the columns of the matrix $(T^{-1})^\top$ are linearly independent. The condition $\operatorname{span}\langle X^\top C^\top\rangle \subset \operatorname{span}\langle X^\top \Sigma\rangle$ is satisfied if and only if $\lambda_i \neq 0$ for all i such that $\mu_i \neq 0$ (that is $\nu_i < \infty$, $i = 1, \dots, d$, where notation $\nu_i = \lambda_i/\nu_i$ comes from Theorem 6.2). Thus, due to Proposition 6.3,

$$\bigl(X^\top \Sigma X\bigr)^\dagger = T\mathrm{M}^\dagger T^\top.$$

Construct the chain of equalities:

$$\begin{aligned}
&\min_{\substack{\Delta\,(I-P_\Sigma)=0\\ (C-\Delta)X=0}} \lambda_{k+m-d}\bigl(\Delta\Sigma^\dagger\Delta^\top\bigr)\\
&\quad\overset{\text{(a)}}{=} \lambda_{k+m-d}\bigl(CX\bigl(X^\top\Sigma X\bigr)^\dagger X^\top C^\top\bigr) = \lambda_{k+m-d}\bigl(CX\,T\mathrm{M}^\dagger T^\top\,X^\top C^\top\bigr)\\
&\quad\overset{\text{(b)}}{=} \lambda_k\bigl(\mathrm{M}^\dagger T^\top X^\top C^\top CXT\bigr) = \lambda_k\bigl(\mathrm{M}^\dagger \Lambda\bigr) = \nu_k\\
&\quad\overset{\text{(c)}}{=} \min\bigl\{\lambda \geq 0 : \text{“}\exists V_1{\subset}\mathbb{R}^d,\ \dim V_1{=}k : \bigl(X^\top C^\top CX - \lambda X^\top\Sigma X\bigr)|_{V_1} \leq 0\text{”}\bigr\}\\
&\quad\overset{\text{(d)}}{=} \min\bigl\{\lambda \geq 0 : \text{“}\exists V{\subset}\operatorname{span}\langle X\rangle,\ \dim V{=}k : \bigl(C^\top C - \lambda\Sigma\bigr)|_{V} \leq 0\text{”}\bigr\}.
\end{aligned}$$

Equality (a) follows from 7.2 because the matrix $CX(X^\top\Sigma X)^\dagger X^\top C^\top$ is the least value of the expression $\Delta\Sigma^\dagger\Delta^\top$ with constraints $(I - P_\Sigma)\Delta^\top = 0$ and $(C - \Delta)X = 0$.

Equality (b) follows from the relation between characteristic polynomials of two products of two rectangular matrices:

$$\chi_{CXT\,\mathrm{M}^\dagger T^\top X^\top C^\top}(\lambda) = (-\lambda)^{m-d}\chi_{\mathrm{M}^\dagger T^\top X^\top C^\top\,CXT}(\lambda)$$

because CXT is an $m\times d$ matrix and $\mathrm{M}^\dagger T^\top X^\top C^\top$ is a $d\times m$ matrix. Thus, the matrix $CXT\,\mathrm{M}^\dagger T^\top X^\top C^\top$ has all the eigenvalues of the matrix $\mathrm{M}^\dagger T^\top X^\top C^\top \times CXT = \mathrm{M}^\dagger\Lambda$ and, besides them, the eigenvalue 0 of multiplicity $m - d$. All these eigenvalues are nonnegative.

Equality (c) holds true due to Lemma 7.1.

Since the columns of the matrix X are linearly independent, there is a one-to-one correspondence between subspaces of $\operatorname{span}\langle X\rangle$ and of $\mathbb{R}^d$: if V is a subspace of $\operatorname{span}\langle X\rangle$, then there exists a unique subspace $V_1 \subset \mathbb{R}^d$, and for those V and V_1,

- $\dim V = \dim V_1$;
- the restriction of the quadratic form $C^\top C - \lambda\Sigma$ to the subspace V is negative semidefinite if and only if the restriction of the quadratic form $X^\top C^\top CX - \lambda X^\top\Sigma X$ to the subspace V_1 is negative semidefinite.

Hence, equality (d) holds true.

Equation (34) is proved. As to Remark 7.4-1, the minimum in the left-hand side of (34) is attained for $\Delta = \Delta_{\mathrm{pm}}$. The minimum in the right-hand side of (34) is attained if the subspace V is a linear span of k columns of the matrix XT that correspond to the k least ν_i's. □

Proof of Proposition 7.5. By Lemma 7.1 and Proposition 7.4, the inequality (37) is equivalent to the obvious inequality

$$\begin{aligned}&\min\bigl\{\lambda \geq 0 : \text{“}\exists V \subset \operatorname{span}\langle X\rangle,\ \dim V = k : \bigl(C^\top C - \lambda \Sigma\bigr)|_V \leq 0\text{”}\bigr\}\\ &\quad \geq \min\bigl\{\lambda \geq 0 \mid \text{“}\exists V,\ \dim V = k : (A - \lambda B)|_V \leq 0\text{”}\bigr\}.\end{aligned}$$

From the proof it follows that if $\nu_d = \infty$, then for any $(n+d) \times d$ matrix X of rank d the constraints in (28) are not compatible.

Now prove that if $\nu_d < \infty$ and $X = [u_1, u_2, \ldots, u_d]$, then the inequality in Proposition 7.5 becomes an equality. Indeed, then the constraints in (28) are compatible because they are satisfied for $\Delta = CTDT^{-1}$, where

$$D = \operatorname{diag}(d_1, d_2, \ldots, d_{d+n}),$$
$$d_k = \begin{cases} 1 & \text{if } \mu_k > 0 \text{ and } k \leq d, \\ 0 & \text{if } \mu_k = 0 \text{ or } k > d. \end{cases}$$

By Proposition 7.2

$$\begin{aligned}\min_{\substack{\Delta(I-P_\Sigma)=0\\(C-\Delta)X=0}} \lambda_{k+m-d}\bigl(\Delta\Sigma^\dagger\Delta^\top\bigr) &= \lambda_{k+m-d}\bigl(CX\bigl(X^\top\Sigma X\bigr)^\dagger X^\top C^\top\bigr)\\ &= \lambda_k\bigl(\bigl(X^\top\Sigma X\bigr)^\dagger X^\top C^\top CX\bigr)\\ &= \lambda_k\bigl(\mathrm{M}_d^\dagger \Lambda_d\bigr) = \nu_k,\end{aligned}$$

where $\mathrm{M}_d = \operatorname{diag}(\mu_1, \ldots, \mu_d)$ and $\Lambda_d = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ are principal submatrices of the matrices M and Λ, respectively. □

Proof of Proposition 7.6. For every matrix Δ that satisfies the constraints $(I - P_\Sigma)\Delta = 0$ and $\operatorname{rk}(C - \Delta) \leq n$, there exists an $(n+d) \times d$ matrix X of rank d such that $(C - \Delta)X = 0$. Assuming that such Δ exists, we get $\nu < +\infty$ because the equalities $\nu = +\infty$, $(I - P_\Sigma)\Delta = 0$, $\operatorname{rk} X = d$, and $(C - \Delta)X = 0$ cannot hold simultaneously.

We have

$$\begin{aligned}\bigl\|\Delta\,\bigl(\Sigma^{1/2}\bigr)^\dagger\bigr\|_F^2 &= \operatorname{tr}\bigl(\Delta\Sigma^\dagger\Delta^\top\bigr) = \sum_{i=1}^{m} \lambda_i\bigl(\Delta\Sigma^\dagger\Delta^\top\bigr)\\ &= \sum_{i=1}^{m-d} \lambda_i\bigl(\Delta\Sigma^\dagger\Delta^\top\bigr) + \sum_{k=1}^{d} \lambda_{k+m-d}\bigl(\Delta(\Sigma)^\dagger\Delta^\top\bigr)\\ &\geq 0 + \sum_{k=1}^{d} \nu_k, \qquad (73)\end{aligned}$$

where the inequalities hold true due to positive semidefiniteness of Σ and due to Proposition 7.5.

If $\nu_d = \infty$, than the constraints $\Delta\,(I - P_\Sigma) = 0$ and $\operatorname{rk}(C - \Delta) \leq n$ are not compatible. Otherwise, the equality in (73) is attained for $\Delta = \Delta_{\mathrm{em}} := CX(X^\top\Sigma X)^\dagger \times$

$X^\top \Sigma$, where the matrix X consists of first d rows of the matrix T, where T comes from decomposition (35).

Thus, if the constraints in (7) are compatible, then the minimum is equal to $(\sum_{k=1}^d \nu_k)^{1/2}$ and is attained at Δ_{em}. Otherwise, if the constraints are incompatible, then by contraposition to the second statement of Proposition 7.5 $\nu_d = +\infty$ and $(\sum_{k=1}^d \nu_k)^{1/2} = +\infty$.

If the minimum in (7) is attained at Δ, then the inequality (73) becomes an equality, whence

$$\lambda_i\big(\Delta\Sigma^\dagger\Delta^\top\big) = 0, \quad i = 1, \ldots, m-d; \tag{74}$$

$$\lambda_{k+m-d}\big(\Delta\Sigma^\dagger\Delta^\top\big) = \nu_k, \quad k = 1, \ldots, d; \tag{75}$$

in particular,

$$\lambda_{\max}\big(\Delta\Sigma^\dagger\Delta^\top\big) = \nu_d.$$

Remember that ν_d is the minimum value in (11). Thus, the minimum in (11) is attained at Δ, although it may be also attained elsewhere. □

Proof of Proposition 7.7. *1.* The monotonicity follows from results of [14]. The unitarily invariant norm is a symmetric gauge function of the singular values, and the symmetric gauge function is monotonous in non-negative inputs (see [14, ineq. (2.5)]).

2. Let $\sigma_1(M_1) < \sigma_1(M_2)$ and $\sigma_i(M_1) \le \sigma_i(M_2)$ for all $i = 2, \ldots, \min(m, n)$. Then for all $k = 1, \ldots, \min(m, n)$

$$\sum_{i=1}^k \sigma_i(M_1) \le \frac{\sigma_1(M_1) + \sigma_2(M_1) + \cdots + \sigma_{\min(m,n)}(M_1)}{\sigma_1(M_2) + \sigma_2(M_1) + \cdots + \sigma_{\min(m,n)}(M_1)} \sum_{i=1}^k \sigma_i(M_2).$$

Due to Ky Fan [3, Theorem 4] or [14, Theorem 1], this implies that

$$\|M_1\|_{\mathrm{U}} \le \frac{\sigma_1(M_1) + \sigma_2(M_1) + \cdots + \sigma_{\min(m,n)}(M_1)}{\sigma_1(M_2) + \sigma_2(M_1) + \cdots + \sigma_{\min(m,n)}(M_1)} \|M_2\|_{\mathrm{U}}.$$

Since

$$0 \le \frac{\sigma_1(M_1) + \sigma_2(M_1) + \cdots + \sigma_{\min(m,n)}(M_1)}{\sigma_1(M_2) + \sigma_2(M_1) + \cdots + \sigma_{\min(m,n)}(M_1)} < 1 \quad \text{and} \quad \|M_2\|_{\mathrm{U}} > 0,$$

$\|M_1\|_{\mathrm{U}} < \|M_2\|_{\mathrm{U}}$. □

Proof of Proposition 7.8. Notice that the optimization problems (7), (11), and (12) have the same constraints. If the constraints are compatible, then the minimum in (7) is attained for $\Delta = \Delta_{\mathrm{em}} := CX(X^\top\Sigma X)^\dagger X^\top\Sigma$.

1. Let $\Delta_{\min\,(7)}$ minimize (7), and let Δ_{feas} satisfy the constraints. Then, by Proposition 7.5 and eqn. (75),

$$\lambda_{k+m-d}\big(\Delta_{\min\,(7)}\Sigma^\dagger\Delta_{\min\,(7)}^\top\big) = \nu_k \le \lambda_{k+m-d}\big(\Delta_{\mathrm{feas}}\Sigma^\dagger\Delta_{\mathrm{feas}}^\top\big), \quad k = 1, \ldots, d;$$

$$\sigma_{d+1-k}\big(\Delta_{\min\,(7)}\big(\Sigma^{1/2}\big)^\dagger\big) \le \sigma_{d+1-k}\big(\Delta_{\mathrm{feas}}\big(\Sigma^{1/2}\big)^\dagger\big),$$

$$\begin{gathered}k = \max(1, d+1-m), \dots, d; \\ \sigma_j\big(\Delta_{\min(7)}\big(\Sigma^{1/2}\big)^\dagger\big) \le \sigma_j\big(\Delta_{\text{feas}}\big(\Sigma^{1/2}\big)^\dagger\big), \quad j = 1, \dots, \min(d, m);\end{gathered}$$

by eqn. (74)

$$\begin{aligned}\lambda_i\big(\Delta_{\min(7)}\Sigma^\dagger\Delta_{\min(7)}^\top\big) &= 0, \quad i = 1, \dots, m-d, \\ \sigma_{m+1-i}\big(\Delta_{\min(7)}\big(\Sigma^{1/2}\big)^\dagger\big) &= 0 \le \sigma_{m+1-i}\big(\Delta_{\text{feas}}\big(\Sigma^{1/2}\big)^\dagger\big), \quad i \le m-d; \\ \sigma_j\big(\Delta_{\min(7)}\big(\Sigma^{1/2}\big)^\dagger\big) &= 0 \le \sigma_j\big(\Delta_{\text{feas}}\big(\Sigma^{1/2}\big)^\dagger\big), \quad d+1 \le j \le \min(m,\ n+d).\end{aligned}$$

Thus

$$\sigma_j\big(\Delta_{\min(7)}\big(\Sigma^{1/2}\big)^\dagger\big) \le \sigma_j\big(\Delta_{\text{feas}}\big(\Sigma^{1/2}\big)^\dagger\big) \quad \text{for all } j \le \min(m, n+d), \tag{76}$$

whence by Proposition 7.7 $\|\Delta_{\min(7)}(\Sigma^{1/2})^\dagger\|_{\mathrm{U}} \le \|\Delta_{\text{feas}}(\Sigma^{1/2})^\dagger\|_{\mathrm{U}}$. Thus $\Delta_{\min(7)}$ indeed minimizes (12).

2. Let $\Delta_{\min(12)}$ minimize (12), so the constraints are compatible. Then Δ_{em} minimizes both (7) and (11), see Proposition 7.6. Thus,

$$\big\|\Delta_{\min(12)}\big(\Sigma^{1/2}\big)^\dagger\big\|_{\mathrm{U}} \le \big\|\Delta_{\text{em}}\big(\Sigma^{1/2}\big)^\dagger\big\|_{\mathrm{U}},$$

and by (76)

$$\sigma_j\big(\Delta_{\text{em}}\big(\Sigma^{1/2}\big)^\dagger\big) \le \sigma_j\big(\Delta_{\min(12)}\big(\Sigma^{1/2}\big)^\dagger\big) \quad \text{for all } j \le \min(m, n+d).$$

Then by Proposition 7.7 (contraposition to part 2)

$$\begin{gathered}\sigma_1\big(\Delta_{\text{em}}\big(\Sigma^{1/2}\big)^\dagger\big) = \sigma_1\big(\Delta_{\min(12)}\big(\Sigma^{1/2}\big)^\dagger\big), \\ \min_{\substack{\Delta(I-P_\Sigma)=0 \\ \operatorname{rk}(C-\Delta)\le n}}\big(\Delta\Sigma^\dagger\Delta^\top\big) = \lambda_{\max}\big(\Delta_{\text{em}}\Sigma^\dagger\Delta_{\text{em}}^\top\big) = \lambda_{\max}\big(\Delta_{\min(12)}\Sigma^\dagger\Delta_{\min(12)}^\top\big).\end{gathered}$$

Thus $\Delta_{\min(12)}$ indeed minimizes (11). □

Proof of Proposition 7.9. We can assume that $\mu_i \in \{0, 1\}$ in (35).

The set of matrices Δ that satisfy (8) depends only on $\operatorname{span}\langle\widehat{X}_{\text{ext}}\rangle$ and does not change after linear transformations of columns of $\widehat{X}_{\text{ext}}$.

By linear transformations of the columns, the matrix $T^{-1}\widehat{X}_{\text{ext}}$ can be transformed to the reduced column echelon form. Thus, there exists such an $(n+d) \times d$ matrix T_5 in the column echelon form that

$$\operatorname{span}\langle\widehat{X}_{\text{ext}}\rangle = \operatorname{span}\langle T T_5\rangle.$$

Notice that $\operatorname{rk} T_5 = \operatorname{rk}\widehat{X}_{\text{ext}} = d$.

Denote by d_* and d^* the first and the last of the indices i such that $\nu_i = \nu_d$. Then

$$\begin{aligned}\nu_{d_*-1} &< \nu_{d_*} \quad \text{if } d_* \ge 2; \\ \nu_{d_*} &= \dots = \nu_d = \dots = \nu_{d^*}; \\ \nu_{d^*} &< \nu_{d^*+1} \quad \text{if } d^* < n+d.\end{aligned}$$

Necessity. Let Δ be a point where the constrained minimum in (7) is attained. Then equalities (74)–(75) from the proof of Proposition 7.6 hold true. Thus, due to Propositions 7.4 and 7.5, for all $k = 1, \dots, d$

$$\min\bigl\{\lambda \ge 0 : \text{“}\exists V \subset \operatorname{span}\langle \widehat{X}_{\text{ext}}\rangle,\ \dim V = k : \bigl(C^\top C - \lambda\Sigma\bigr)|_V \le 0\text{”}\bigr\} = \nu_k.$$

According to 7.4-1, we can construct a stack of subspaces

$$V_1 \subset V_2 \subset \cdots \subset V_d = \operatorname{span}\langle \widehat{X}_{\text{ext}}\rangle,$$

such that $\dim V_k = k$ and the restriction of the quadratic form $C^\top C - \nu_k\Sigma$ to the subspace V_k is negative semidefinite, for all $k \le d$.

Now, prove that

$$\operatorname{span}\langle u_i : \nu_i < \nu_d\rangle \subset \operatorname{span}\langle \widehat{X}_{\text{ext}}\rangle. \tag{77}$$

Suppose the contrary: $\operatorname{span}\langle u_i : \nu_i < \nu_d\rangle \not\subset \operatorname{span}\langle \widehat{X}_{\text{ext}}\rangle$. Then there exists $i < d_*$ such that $u_i \notin \operatorname{span}\langle \widehat{X}_{\text{ext}}\rangle$, and, as a consequence, $u_i \notin V_{\max\{j\,:\,\nu_j \le \nu_i\}}$. Find the least k such that $u_k \notin V_{\max\{j\,:\,\nu_j \le \nu_k\}}$. Let k_* and k^* denote the first and the last indices i such that $\nu_i = \nu_k$. Then $1 \le k_* \le k \le k^* < d_* \le d \le d^*$ and $u_k \notin V_{k^*}$.

Since $\operatorname{span}\langle u_1, \dots, u_{k_*-1}\rangle \subset V_{k_*-1} \subset V_{k^*}$,

$$\begin{aligned}\dim\bigl(V_{k^*} \cap \operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle\bigr) &= \dim\bigl(V_{k^*}/\operatorname{span}\langle u_1, \dots, u_{k_*-1}\rangle\bigr)\\ &= \dim V_{k^*} - (k_* - 1) = k^* - k_* + 1.\end{aligned}$$

Since $u_k \notin V_{k^*}$, $u_k \notin V_{k^*} \cap \operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle$,

$$\dim\operatorname{span}\bigl\langle V_{k^*} \cap \operatorname{span}\langle u_{k_*}, \dots u_{n+d}\rangle,\ u_k\bigr\rangle = k^* - k_* + 2.$$

Now, consider the $(n + d - k_* + 1) \times (n + d - k_* + 1)$ diagonal matrix

$$\begin{aligned}D(\lambda) &:= [u_{k_*}, \dots, u_{n+d}]^\top\bigl(C^\top C - \lambda\Sigma\bigr)[u_{k_*}, \dots, u_{n+d}]\\ &= \operatorname{diag}(\lambda_j - \lambda\mu_j,\ \ j = k_*, \dots, n+d)\end{aligned}$$

for various λ. For $\lambda = \nu_k = \nu_{k_*}$, the inequality $\lambda_j - \nu_k\mu_j \ge 0$ holds true for all $j \ge k_*$, so the matrix $D(\nu_k)$ is positive semidefinite. For $\lambda = \nu_{k^*+1}$, the inequality $\lambda_j - \nu_{k^*+1}\mu_j \le 0$ holds true for all $k_* \le j \le k^* + 1$, so there exists a $k^* - k_* + 2$-dimensional subspace of $\mathbb{R}^{n+d-k_*+1}$ where the quadratic form $D(\nu_{k^*+1})$ is negative semidefinite. For $\lambda < \nu_{k^*+1}$, the inequality $\lambda_j - \lambda\mu_j > 0$ holds true for all $k^* + 1 \le j \le n + d$. Therefore, there exists an $n + d - k^*$-dimensional subspace of $\mathbb{R}^{n+d-k_*+1}$ where the quadratic form $D(\lambda)$ is positive definite. According to the proof of Sylvester's law of inertia, there is no subspace of dimension $k^* - k_* + 2 = (n + d - k_* + 1) - (n + d - k^*) + 1$ where the quadratic form $D(\lambda)$ is negative semidefinite. Thus, ν_{k^*+1} is the least number such that there exists a $k^* - k_* + 2$-dimensional subspace where the quadratic form $D(\lambda)$ is negative semidefinite.

Similarly to the chain of equalities in the proof of Proposition 7.4,

$$\begin{aligned}\nu_{k^*+1} &= \min\bigl\{\lambda \ge 0 : \text{“}\exists V_1,\ \dim V_1 = k^* - k_* + 2\ :\ D(\lambda)|_{V_1} \le 0\text{”}\bigr\}\\ &= \min\bigl\{\lambda \ge 0 : \text{“}\exists V_1,\ \dim V_1 = k^* - k_* + 2\ :\end{aligned}$$

$$\begin{aligned}&[u_{k_*}, \dots, u_{n+d}]^\top \bigl(C^\top C - \lambda \Sigma\bigr)[u_{k_*}, \dots, u_{n+d}]|_{V_1} \le 0\text{”}\bigr\} \\ &= \min\bigl\{\lambda \ge 0 : \text{“}\exists V_1,\ V \subset \operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle,\ \dim V = k^* - k_* + 2\ : \\ &\quad \bigl(C^\top C - \lambda \Sigma\bigr)|_V \le 0\text{”}\bigr\} \end{aligned} \tag{78}$$

The restriction of the quadratic form $C^\top C - \nu_k \Sigma$ to the subspace $\operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle$ is positive semidefinite because $[u_{k_*}, \dots, u_{n+d}]^\top (C^\top C - \nu_k \Sigma) \times [u_{k_*}, \dots, u_{n+d}] = D(\nu_k)$ is a positive semidefinite diagonal matrix. Then

$$\begin{aligned}&\bigl\{v \in \operatorname{span}\langle u_{k_*}, \dots u_{n+d}\rangle : v^\top \bigl(C^\top C - \nu_k \Sigma\bigr) v \le 0\bigr\} \\ &\quad = \bigl\{v \in \operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle : \bigl(C^\top C - \nu_k \Sigma\bigr) v = 0\bigr\}\end{aligned} \tag{79}$$

is a linear subspace. Since this subspace contains the subspace $V_k \cap \operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle$ (as the quadratic form $C^\top C - \nu_k \Sigma$ is negative semidefinite on V_k) and the vector u_k (as $u_k \in \operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle$ and $u_k^\top (C^\top C - \nu_k \Sigma) u_k = \lambda_k - \nu_k \mu_k = 0$), it contains $\operatorname{span}\langle V_{k^*} \cap \operatorname{span}\langle u_{k_*}, \dots, u_{n+d}\rangle,\ u_k\rangle$. But, as $\nu_k < \nu_{k^*+1}$, this contradicts (78).

Now, prove that

$$\operatorname{span}\langle \widehat{X}_{\mathrm{ext}}\rangle \subset \operatorname{span}\langle u_i : \nu_i \le \nu_d\rangle. \tag{80}$$

Due to (77),

$$\operatorname{span}\langle \widehat{X}_{\mathrm{ext}}\rangle = \operatorname{span}\bigl\langle \operatorname{span}\langle \widehat{X}_{\mathrm{ext}}\rangle \cap \operatorname{span}\langle u_{d_*}, \dots, u_{n+d}\rangle,\ u_1, \dots, u_{d_*-1}\bigr\rangle.$$

Hence, to prove (80), it is enough to show that

$$\operatorname{span}\langle \widehat{X}_{\mathrm{ext}}\rangle \cap \operatorname{span}\langle u_{d_*}, \dots, \nu_{n+d}\rangle \subset \operatorname{span}\langle u_{d_*}, \dots, \nu_{d^*}\rangle. \tag{81}$$

The restriction of the quadratic form $C^\top C - \nu_d \Sigma$ to the subspace $\operatorname{span}\langle u_{d_*}, \dots, u_{n+d}\rangle$ is positive semidefinite. Hence

$$\begin{aligned}&\bigl\{v \in \operatorname{span}\langle u_{d_*}, \dots u_{n+d}\rangle : v^\top \bigl(C^\top C - \nu_d \Sigma\bigr) v \le 0\bigr\} \\ &\quad = \bigl\{v \in \operatorname{span}\langle u_{d_*}, \dots u_{n+d}\rangle : v^\top \bigl(C^\top C - \nu_d \Sigma\bigr) v = 0\bigr\}\end{aligned} \tag{82}$$

is a linear subspace (see equation (79)). This subspace contains the subspaces $\operatorname{span}\langle \widehat{X}_{\mathrm{ext}}\rangle \cap \operatorname{span}\langle u_{d_*}, \dots, \nu_{n+d}\rangle$ and $\operatorname{span}\langle u_{d_*}, \dots, \nu_{d^*}\rangle$. Denote the dimension of the subspace (82):

$$d_2 = \dim\bigl\{v \in \operatorname{span}\langle u_{d_*}, \dots u_{n+d}\rangle : v^\top \bigl(C^\top C - \nu_d \Sigma\bigr) v = 0\bigr\}.$$

If (81) does not hold, then $d_2 > d^* - d_* + 1$; $d_2 \ge d^* - d_* + 2$. Then

$$\exists V \subset \operatorname{span}\langle u_{d_*}, \dots u_{n+d}\rangle,\ \dim V = d_2\ :\ \bigl(C^\top C - \nu_d \Sigma\bigr)|_V \le 0$$

(as an instance of such a subspace V, we can take the one defined in (82)). Then, taking a $d^* - d_* + 2$-dimensional subspace of V, we get

$$\exists V \subset \operatorname{span}\langle u_{d_*}, \dots u_{n+d}\rangle,\ \dim V = d^* - d_* + 2\ :\ \bigl(C^\top C - \nu_d \Sigma\bigr)|_V \le 0.$$

Due to (78) (for $k = d$), $\nu_{d^*+1} \le \nu_d$, which does not hold true.

Assuming the contrary to (81), we got a contradiction. Hence, (81) and (80) hold true.

Sufficiency. Remember that $T = [u_1, \dots, u_{n+d}]$ is an $(n + d) \times (n + d)$ matrix of generalized eigenvectors of the matrix pencil $\langle C^\top C,\ \Sigma\rangle$, and respective generalized eigenvalues are arranged in ascending order. By means of linear operations of the columns, the matrix $T^{-1}\widehat{X}_{\mathrm{ext}}$ can be transformed into the reduced column echelon form. In other words, there exists such an $n \times n$ nonsingular matrix T_8, that the $(n + d) \times n$ matrix

$$T_5 = T^{-1}\widehat{X}_{\mathrm{ext}}T_8 \tag{83}$$

is in the reduced column echelon form. The equality (83) implies that

$$\operatorname{span}\langle\widehat{X}_{\mathrm{ext}}\rangle = \operatorname{span}\langle T T_5\rangle. \tag{84}$$

If condition (37) holds, then in representation (84) the matrix T_5 has the following block structure

$$T_5 = \begin{array}{|c|c|}\hline I_{d^*-1} & 0_{(d^*-1)\times(d-d^*+1)} \\ \hline 0_{(d^*-d_*+1)\times(d^*-1)} & T_{61} \\ \hline \multicolumn{2}{|c|}{0_{(n-d^*)\times d}} \\ \hline \end{array}\,,$$

where T_{61} is a $(d^* - d_* + 1) \times (d - d_* + 1)$ reduced column echelon matrix. (Any of the blocks except T_{61} may be an "empty matrix".)

Since the columns of T_5 are linearly independent, the columns of T_{61} are linearly independent as well. Hence the matrix T_{61} may be appended with columns such that the resulting matrix $T_6 = [T_{61}, T_{62}]$ is nonsingular. Perform the Gram–Schmidt orthogonalization of columns of the matrix T_6 by constructing such an upper-triangular matrix

$$T_7 = \begin{pmatrix} T_{71} & T_{72} \\ 0 & T_{74} \end{pmatrix} = \begin{array}{|c|c|}\hline T_{71} & T_{72} \\ \hline 0_{(d^*-d)\times(d-d_*+1)} & T_{74} \\ \hline \end{array}$$

that $T_7^\top T_6^\top T_6 T_7 = I_{d^*-d_*+1}$.

Change the basis in the simultaneous diagonalization of the matrices $C^\top C$ and Σ. Denote

$$T_{\mathrm{new}} = \bigl[u_1, \dots u_{d_*-1}, [u_{d_*}, \dots u_{d^*}]T_6T_7, u_{d^*+1}, \dots u_{n+d}\bigr].$$

If $\nu_d > 0$, the equation (35) with T_{new} substituted for T holds true, since

$$T_{\mathrm{new}}^\top C^\top C T_{\mathrm{new}} = \Lambda, \qquad T_{\mathrm{new}}^\top \Sigma T_{\mathrm{new}} = \mathrm{M}.$$

(Here we use that $\lambda_{d_*} = \cdots = \lambda_{d^*}$, $\mu_{d_*} = \cdots = \mu_{d^*}$. If $\nu_d = 0$, then the latter equation may or may not hold true.) The subspace

$$\begin{aligned}\operatorname{span}\langle\widehat{X}_{\mathrm{ext}}\rangle = \operatorname{span}\langle T T_5\rangle &= \operatorname{span}\bigl\langle u_1, \dots u_{d_*-1}, [u_{d_*}, \dots u_{d^*}]T_{61}\bigr\rangle \\ &= \operatorname{span}\bigl\langle u_1, \dots u_{d_*-1}, [u_{d_*}, \dots u_{d^*}]T_{61}T_{71}\bigr\rangle\end{aligned}$$

is spanned by the first d columns of the matrix T_{new}.

It can be easily verified that $\operatorname{span}\langle\widehat{X}_{\mathrm{ext}}^\top C^\top\rangle = \operatorname{span}\langle T_8^\top T_5^\top \Lambda\rangle$ and $\operatorname{span}\langle\widehat{X}_{\mathrm{ext}}^\top \Sigma\rangle = \operatorname{span}\langle T_8^\top T_5^\top \mathrm{M}\rangle$. The condition $\operatorname{span}\langle\widehat{X}_{\mathrm{ext}}^\top C^\top\rangle \subset \operatorname{span}\langle\widehat{X}_{\mathrm{ext}}^\top \Sigma\rangle$ holds true if (and only if) $\nu_d < \infty$. Thus, due to Proposition 7.2, if the condition $\nu_d < \infty$ holds true, then the constraints $\Delta\,(I - P_\Sigma) = 0$ and $(C - \Delta)\widehat{X}_{\text{ext}} = 0$ are compatible.

Let Δ_{pm} be a common point of minimum in

$$\lambda_{k+m-d}\big(\Delta_{\text{pm}}\Sigma^\dagger\Delta_{\text{pm}}^\top\big) = \min_{\substack{\Delta(I-P_\Sigma)=0\\ (C-\Delta)\widehat{X}_{\text{ext}}=0}} \lambda_{k+m-d}\big(\Delta\Sigma^\dagger\Delta^\top\big)$$

for all $k = 1, \ldots, d$, such that $\Delta_{\text{pm}}\,(I - P_\Sigma) = 0$ and $(C - \Delta_{\text{pm}})\widehat{X}_{\text{ext}} = 0$; such Δ_{pm} exists due to Remark 7.4-1. By Proposition 7.5,

$$\lambda_{k+m-d}\big(\Delta_{\text{pm}}\Sigma^\dagger\Delta_{\text{pm}}^\top\big) = \nu_k, \quad k = 1, \ldots, d,$$

and, from the proof of Preposition 7.6,

$$\lambda_i\big(\Delta_{\text{pm}}\Sigma^\dagger\Delta_{\text{pm}}^\top\big) = 0, \quad i = 1, \ldots, m-d.$$

The minimum in (7) is attained at $\Delta = \Delta_{\text{pm}}$.

The case $\nu_d = 0$ is trivial: then (37) imply that $C\widehat{X}_{\text{ext}} = 0$. Then $\Delta = 0$ satisfies the constraints $\Delta\,(I - P_\Sigma) = 0$ and $(C - \Delta)\widehat{X}_{\text{ext}} = 0$ and minimizes the criterion function in (7). □

Proof of Proposition 7.10. Remember that if $\nu_d < \infty$, then the constraints in (11) are compatible, and the minimum is attained and is equal to ν_d; see Proposition 7.5. Otherwise, if $\nu_d = \infty$, then the constraints in (11) are incompatible.

Transform the expression for the functional (38):

$$\begin{aligned} Q_1(X) &:= \lambda_{\max}\big((X^\top\Sigma X)^{-1}X^\top C^\top C X\big) \\ &= \lambda_{\max}\big(CX(X^\top\Sigma X)^{-1}X^\top C^\top\big) \\ &= \min_{\Delta_1\in\mathbb{R}^{m\times(n+d)}\,:\,\Delta_1(I-P_\Sigma)=0,\ (C-\Delta_1)X=0} \lambda_{\max}\big(\Delta_1\Sigma^\dagger\Delta_1^\top\big). \end{aligned} \tag{85}$$

Here we used the rule how eigenvalues of the matrix product change when the matrices are swapped, and we also used Propositions 7.2 and 7.3. By Proposition 7.5, $Q_1(X) \geq \nu_d$.

If the minimum in (11) & (8) is attained (say at some point $(\Delta, \widehat{X}_{\text{ext}})$), then the constraints in the right-hand side of (85) are compatible for $X = \widehat{X}_{\text{ext}}$ (particularly, Δ is a matrix that satisfies the constraints). Then by Proposition 7.3 the matrix $\widehat{X}_{\text{ext}}^\top\Sigma\widehat{X}_{\text{ext}}$ is nonsingular. Thus, for $X = \widehat{X}_{\text{ext}}$, minimum in the right-hand of (85) is attained at $\Delta_1 = \Delta$ (because Δ satisfies stronger constraints of (85) and brings a minimum to the same functional with weaker constraints of (11)).

Hence,

$$\begin{aligned} Q_1(\widehat{X}_{\text{ext}}) &:= \min_{\Delta_1\in\mathbb{R}^{m\times(n+d)}\,:\,\Delta_1(I-P_\Sigma)=0,\ (C-\Delta_1)\widehat{X}_{\text{ext}}=0} \lambda_{\max}\big(\Delta_1\Sigma^\dagger\Delta_1^\top\big) \\ &= \big(\Delta\Sigma^\dagger\Delta^\top\big) = \nu_d, \end{aligned}$$

which is the minimum value of Q_1.

Transform the expression for the functional (39):

$$\lambda_{\max}\big((X^\top \Sigma X)^{-1} X^\top (C^\top C - m\Sigma) X\big)$$
$$= \lambda_{\max}\big((X^\top \Sigma X)^{-1} X^\top (C^\top C) X - m I_{n+d}\big) = Q_1(X) - m.$$

Hence, the functionals (38) and (39) attain their minimal values at the same points. □

9 Conclusion

The linear errors-in-variables model is considered. The errors are assumed to have the same covariance matrix for each observation and to be independent between different observations, however some variables may be observed without errors. Detailed proofs of the consistency theorems for the TLS estimator, which were first stated in [18], are presented.

It is proved that that the final estimator $\widehat{X}$ for explicit-notation regression coefficients (i.e., for X_0 in (1) or (2), and not the estimator $\widehat{X}_{\mathrm{ext}}$ for X^0_{ext} in equation (3), which sets the relationship between the regressors and response variables *implicitly*) is unique, either with high probability or eventually. This means that in the classification used in [8], the TLS problem is of 1st class set $\mathcal{F}_1$ (the solution is unique and "generic"), with high probability or eventually.

As by-product, we get that if in the definition of the estimator the Frobenius norm is replaced by the spectral norm, then the consistency theorems still hold true. The disadvantage of using spectral norm is that the estimator $\widehat{X}$ is not unique then. (The set of solutions to the minimal spectral norm problem contains the set of solutions to the TLS problem. On the other hand, it is possible that the minimal spectral norm problem has solutions, but the TLS problem has not – this is the TLS problem of 1st class set $\mathcal{F}_3$; the probability of this random event tends to 0.)

Results can be generalized to any unitary invariant matrix norm. I do not know whether they hold true for non-invariant norms such as the maximum absolute entry, which is studied in [7].

References

[1] Cheng, C.-L., Van Ness, J.W.: Statistical Regression with Measurement Error. Wiley (2010). MR1719513

[2] de Leeuw, J.: Generalized eigenvalue problems with positive semi-definite matrices. Psychometrika **47**(1), 87–93 (1982). MR0668507. https://doi.org/10.1007/BF02293853

[3] Fan, K.: Maximum properties and inequalities for the eigenvalues of completely continuous operators. Proceedings of the National Academy of Sciences of the USA **37**(11), 760–766 (1951). MR0045952. https://doi.org/10.1073/pnas.37.11.760

[4] Gallo, P.P.: Consistency of regression estimates when some variables are subject to error. Communications in Statistics – Theory and Methods **11**(9), 973–983 (1982). https://doi.org/10.1080/03610928208828287

[5] Gleser, L.J.: Estimation in a multivariate "errors in variables" regression model: Large sample results. The Annals of Statistics **9**(1), 24–44 (1981). MR0600530

[6] Golub, G.H., Hoffman, A., Stewart, G.W.: A generalization of the Eckart–Young–Mirsky matrix approximation theorem. Linear Algebra and its Applications **88–89**(Supplement C), 317–327 (1987). MR0882452. https://doi.org/10.1016/0024-3795(87)90114-5

[7] Hladík, M., Černý, M., Antoch, J.: EIV regression with bounded errors in data: total 'least squares' with Chebyshev norm. Statistical Papers (2017). https://doi.org/10.1007/s00362-017-0939-z

[8] Hnětynková, I., Plešinger, M., Sima, D.M., Strakoš, Z., Van Huffel, S.: The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works. SIAM Journal on Matrix Analysis and Applications **32**(3), 748–770 (2011). MR2825323. https://doi.org/10.1137/100813348

[9] Kukush, A., Markovsky, I., Van Huffel, S.: Consistency of the structured total least squares estimator in a multivariate errors-in-variables model. Journal of Statistical Planning and Inference **133**(2), 315–358 (2005). MR2194481. https://doi.org/10.1016/j.jspi.2003.12.020

[10] Kukush, A., Van Huffel, S.: Consistency of elementwise-weighted total least squares estimator in a multivariate errors-in-variables model $AX = B$. Metrika **59**(1), 75–97 (2004). MR2043433. https://doi.org/10.1007/s001840300272

[11] Marcinkiewicz, J., Zygmund, A.: Sur les fonctions indépendantes. Fundamenta Mathematicae **29**, 60–90 (1937). MR0115885

[12] Markovsky, I., Sima, D.M., Van Huffel, S.: Total least squares methods. Wiley Interdisciplinary Reviews: Computational Statistics **2**(2), 212–217 (2010). https://doi.org/10.1002/wics.65

[13] Markovsky, I., Willems, J.C., Van Huffel, S., De Moor, B.: Exact and Approximate Modeling of Linear Systems: A Behavioral Approach. SIAM, Philadelphia (2006). MR2207544. https://doi.org/10.1137/1.9780898718263

[14] Mirsky, L.: Symmetric gauge functions and unitarily invariant norms. The Quarterly Journal of Mathematics **11**(1), 50–59 (1960). MR0114821. https://doi.org/10.1093/qmath/11.1.50

[15] Newcomb, R.W.: On the simultaneous diagonalization of two semi-definite matrices. Quarterly of Applied Mathematics **19**(2), 144–146 (1961). MR0124336. https://doi.org/10.1090/qam/124336

[16] Petrov, V.V.: Limit Theorems of Probability Theory: Sequences of Independent Random Variables. Clarendon Press, Oxford (1995). MR1353441

[17] Pfanzagl, J.: On the measurability and consistency of minimum contrast estimates. Metrika **14**, 249–272 (1969). https://doi.org/10.1007/BF02613654

[18] Shklyar, S.V.: Conditions for the consistency of the total least squares estimator in an errors-in-variables linear regression model. Theory of Probability and Mathematical Statistics **83**, 175–190 (2011). MR2768857. https://doi.org/10.1090/S0094-9000-2012-00850-8

[19] Stewart, G., Sun, J.-g.: Matrix Perturbation Theory. Academic Press, San Diego (1990). MR1061154

[20] Van Huffel, S., Vandewalle, J.: The Total Least Squares Problem: Computational Aspects and Analysis. SIAM, Philadelphia (1991). MR1118607. https://doi.org/10.1137/1.9781611971002