

Document Version

Final published version

Licence

CC BY

Citation (APA)

Khodakaramzadeh, S., Shafiee, S., de Albuquerque Gleizer, G., & Mohajerin Esfahani, P. (2026). A saddle point algorithm for robust data-driven factor model problems. *Automatica*, 190, Article 113095. <https://doi.org/10.1016/j.automatica.2026.113095>

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership. Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.



Brief Paper

A saddle point algorithm for robust data-driven factor model problems[☆]Shabnam Khodakaramzadeh^{a,*}, Soroosh Shafiee^b, Gabriel de Albuquerque Gleizer^a, Peyman Mohajerin Esfahani^{a,c}^a Delft University of Technology, Delft, The Netherlands^b Cornell University, NY, United States^c University of Toronto, Toronto, Canada

ARTICLE INFO

Article history:

Received 11 June 2025

Received in revised form 5 February 2026

Accepted 10 April 2026

Keywords:

Factor model

Covariance matrix estimation

First-order algorithms

Saddle-point problem

Robust optimization

ABSTRACT

We study the factor model problem, which aims to uncover low-dimensional structures in high-dimensional datasets. Adopting a robust data-driven approach, we formulate the problem as a saddle-point optimization. Our primary contribution is a first-order algorithm that solves this reformulation by leveraging a linear minimization oracle (LMO). We further develop semi-closed form solutions (up to a scalar) for three specific LMOs, corresponding to the Frobenius norm, Kullback–Leibler divergence, and Gelbrich (aka Wasserstein) distance. The analysis includes explicit quantification of these LMOs' regularity conditions, notably the Lipschitz constants of the dual function, which govern the algorithm's convergence performance. Numerical experiments confirm our method's effectiveness in high-dimensional settings, outperforming standard off-the-shelf optimization solvers.

© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The correlation structure among random variables can be expressed using factor analysis in terms of common factors (Khamaru & Mazumder, 2019), i.e., the information in high-dimensional data can be compressed into unobserved factors (Zorzi & Sepulchre, 2015). Mathematically, a high-dimensional point $\xi \in \mathbb{R}^n$ is represented as sum of two independent unobserved parts

$$\xi = \Phi\alpha + \omega, \quad (1)$$

where factor loading matrix $\Phi \in \mathbb{R}^{n \times r}$ is tall ($n \gg r$) and full-rank. The vector $\alpha \in \mathbb{R}^r$, with covariance $\Sigma_\alpha = I$, contains independent latent factors. Hence, $\Phi\alpha$, having interrelated components, specifies the low-dimensional representation of ξ . The idiosyncratic noise ω has independent components (Anderson, 2003; Ciccone et al., 2018b). Such low-dimensional representations have been also studied in principal component analysis (PCA), which is well-suited for scenarios where the data is corrupted by small,

unstructured noise (Bishop, 2006), and also compressed sensing, where the noise is small and unstructured, but r factors are contained in a sparse tall vector α (Donoho, 2006). In contrast, the noise in the factor model has independent components and may be substantial. Originating from the analysis of mental test scores (Anderson, 2003; Spearman, 1904), the factor model has found broad applications across various domains, including control and system identification (Bottegal & Picci, 2014; Falconi et al., 2023; Picci et al., 2023), fault and anomaly detection (Wu & Zhang, 2006; Zhou et al., 2018), econometrics (Engle & Watson, 1981; Forni & Lippi, 2001; van Schuppen, 1986), and statistics (Ciccone et al., 2018b). When the covariance of ξ , denoted by Σ , is available, and assuming both random variables ξ and ω are zero-mean and independent from each other, we can rewrite (1) as

$$\Sigma = L + D, \quad (2)$$

where $L = \Phi\Sigma_\alpha\Phi^\top$ is low rank with $\text{rank}(L) = r$ (i.e., number of factors) and the non-negative diagonal matrix D is the noise covariance. Often, Σ is only observable by a finite dataset $\{\xi_k\}_{k=1}^N$, and is thus approximated with

$$\widehat{\Sigma} = \frac{1}{N} \sum_{k=1}^N (\xi_k - \widehat{\mu})(\xi_k - \widehat{\mu})^\top, \quad \widehat{\mu} = \frac{1}{N} \sum_{k=1}^N \xi_k. \quad (3)$$

[☆] This research is supported by the ERC Starting Grant TRUST-949796 and by the NSF CAREER ECCS-2541066. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Mattia Zorzi under the direction of Editor Alessandro Chiuso.

* Corresponding author.

E-mail addresses: S.Khodakaramzadeh@tudelft.nl (S. Khodakaramzadeh), shafiee@cornell.edu (S. Shafiee), G.Gleizer@tudelft.nl (G. de Albuquerque Gleizer), P.MohajerinEsfahani@utoronto.ca (P. Mohajerin Esfahani).

One can consider $\widehat{\mu} = 0$, under the assumption that ξ is zero-mean. To robustify to this approximation error, a family of covariance matrices is considered around $\widehat{\Sigma}$ as

$$\mathbb{B}_\varepsilon^d(\widehat{\Sigma}) := \{\Sigma \succeq 0 : d(\Sigma, \widehat{\Sigma}) \leq \varepsilon\}, \quad (4)$$

where d is a generic distance function in the space of matrices, and ε is the radius of the set. Therefore, our *robust data-driven factor model* problem is formulated as

$$J^* := \min_{L,D} \text{Tr}(L) \quad \text{s.t.} \quad L \in \mathbb{S}_+, D \in \mathbb{D}_+, L + D \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma}) \quad (5)$$

where \mathbb{S}_+ is the cone of positive semidefinite (PSD) matrices, \mathbb{D}_+ is the cone of non-negative diagonal matrices. The last constraint is equivalent to $d(L + D, \widehat{\Sigma}) \leq \varepsilon$. The trace operator is a standard convexification for the rank function (Recht et al., 2010), i.e., the objective is to find the *least number* of factors explaining the data. The hyperparameter ε depends on the precision of the approximation $\widehat{\Sigma}$.

Traditionally, the factor model problem was addressed assuming that $\widehat{\Sigma}$ is an accurate estimate of Σ (i.e., $\varepsilon = 0$) (Bishop, 2006; McLachlan & Krishnan, 2007). However, recent research accounts for uncertainty in $\widehat{\Sigma}$, (i.e., $\varepsilon > 0$). For example, Ciccone et al. (2018a) proposes a coordinate descent-type algorithm to minimize $\|\widehat{\Sigma} - L - D\|_F^2$, while (Bertsimas et al., 2017) develops an algorithm based on conditional gradient method targeting $\|\widehat{\Sigma} - L - D\|_q^q$. The robust factor model problem (5) and its dynamic counterpart, limited to special cases where d is the Kullback–Leibler (KL) and Itakura–Saito divergences, are studied in Ciccone et al. (2018b) and Falconi et al. (2023), respectively. However, unlike (Bertsimas et al., 2017; Ciccone et al., 2018a, 2018b; Falconi et al., 2023), which are limited to a specific choice of d , we introduce an algorithm for the generic d , which only requires access to the linear minimization oracle (LMO)

$$O(A) := \arg \min_{\Sigma} \{\langle A, \Sigma \rangle : \Sigma \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})\} \quad (6)$$

for any symmetric matrix A .¹ From a computational perspective, this study focuses on developing an efficient algorithm tailored to the semidefinite programming (SDP) (5), for which the commercial solvers (e.g., MOSEK Mosek ApS, 2025) typically rely on second-order methods, and hence do not scale well to large problems. Moreover, General-purpose first-order methods (e.g., SCS O'Donoghue et al., 2016) instead require projection oracles that involve solving quadratic objectives over the feasible set, in contrast to the LMO (6), which only requires a linear objective. In fact, the LMO (6) can admit closed-form solutions, thereby avoiding the need to solve a full SDP. It is also worth noting that the problem (5) involves conic constraints associated with \mathbb{S}_+ and \mathbb{D}_+ , which can make the direct solution computationally demanding.

In this work, we consider the static factor model problem, where the factors remain constant over time; for a dynamic counterpart, where the factors temporal evolution is modeled via state-space equations, see, e.g., Picci et al. (2023).

We summarize the contributions of this work as follows.

- (i) **Saddle point characterization.** For possible conic structural information about the covariance matrix and generic distance functions, we reformulate factor model (5) as a saddle point problem (Proposition 2.1).
- (ii) **First-order algorithm & convergence.** Leveraging the saddle point reformulation and given an LMO, we propose a first-order algorithm with convergence guarantees derived based on standard regularity conditions such as Lipschitz constants (Proposition 2.3). A particular feature of the proposed algorithm is

the linear convergence rate of the projection operator, enabled by Dykstra's projection technique, as opposed to the standard sublinear rate (Proposition 2.6).

(iii) **Special LMOs: closed-form description & Lipschitz constant.**

The choice of distance function influences the LMO and Lipschitz constant of the dual function, both of which play a central role in the proposed algorithm convergence. We derive the closed-form description for the LMO and its respective Lipschitz constant for the special cases of Frobenius norm (Proposition 3.1), KL divergence (Proposition 3.2), and Gelbrich distance (Proposition 3.4). As a special byproduct, we show that the Gelbrich distance is strongly convex with respect to the Frobenius norm, a property of interest for optimization algorithms (Remark 3.6). In comparison with the existing literature, to our best knowledge, this appears to be the first result that also applies to low-rank matrices.

The theoretical results are validated through extensive numerical experiments, demonstrating the performance of the proposed algorithm. To improve the flow and readability, all the technical proofs are relegated to Appendix A. To facilitate reproducibility, we provide an open-source MATLAB library available at https://github.com/skhodakaram/Factor_Model.

Notation. For any symmetric matrix A , its maximum eigenvalue is represented as $\lambda_{\max}(A)$, its trace is denoted by $\text{Tr}(A)$, and the vector $\lambda^{\frac{1}{2}}(A)$ contains the square roots of all its eigenvalues. The diagonal matrix $\text{Diag}(A)$ contains the diagonal entries of A . The inner product of any $A, B \in \mathbb{R}^{n \times m}$ is denoted as $\langle A, B \rangle := \text{Tr}(A^T B)$. The Frobenius and nuclear norms of $A \in \mathbb{R}^{n \times m}$ are denoted as $\|A\|_F := \sqrt{\langle A, A \rangle}$ and $\|A\|_*$, respectively. The Euclidean norm of $a \in \mathbb{R}^n$ is denoted by $\|a\|_2$. The relative interior of a set \mathbb{A} is represented as $\text{rint}(\mathbb{A})$. The operator $\Pi_{\mathbb{A}}[x] := \arg \min_{y \in \mathbb{A}} \|x - y\|_F^2$ denotes the orthogonal projection of the point x onto the set \mathbb{A} . The element-wise inequality between matrices is represented as $A \geq B$, and the semidefinite counterpart as $A \succeq B$. The positive semidefinite (PSD) cone is denoted by \mathbb{S}_+ . The dual cone of a convex cone $\mathbb{C} \in \mathbb{R}^n$ is represented as $\mathbb{C}^* := \{x \in \mathbb{R}^n : \langle x, y \rangle \geq 0, \forall y \in \mathbb{C}\}$. The space of all lower triangular matrices in $\mathbb{R}^{n \times n}$ is denoted by \mathbb{L}_n . The normal cone of a set \mathbb{A} at x is defined by $\mathcal{N}_{\mathbb{A}}(x) := \{v : \langle v, y - x \rangle \leq 0, \forall y \in \mathbb{A}\}$. The indicator function over the set \mathbb{A} is denoted as $\mathbb{1}_{\mathbb{A}}$.

2. Saddle point reformulation and first-order algorithm

2.1. Saddle point characterization

The first result of this paper is a saddle point (max–min) reformulation of the factor model problem (5). This reformulation paves the way for optimization algorithms, especially considering the availability of the LMO (6).

Proposition 2.1 (Saddle Point Reformulation). *The optimal value J^* of the factor model problem in (5) is equivalent to the max–min problem*

$$J^* = \max_{L \in \mathbb{S}_+} \min_{\Sigma \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle L, \Sigma \rangle, \quad (7)$$

$$-L \in \mathbb{D}_+^*$$

where $\mathbb{D}_+^* = \{A : \text{Diag}(A) \geq 0\}$ is the dual cone of \mathbb{D}_+ .

We note that LMO (6) is used to find the solution to the inner minimization of (7) as a function of the decision variable of the outer maximization problem. The max–min problem (7) facilitates development of an optimization algorithm to tackle (5) numerically. The inner minimal value, hereafter referred to as the dual function, is

$$g(A) := \min_{\Sigma \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle A, \Sigma \rangle \quad (8)$$

¹ If the program (6) has multiple optimizers, the oracle $O(A)$ can be an arbitrary selection from the set.

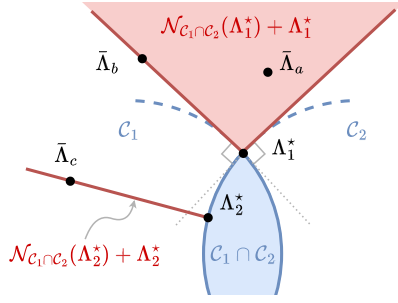


Fig. 1. Depiction of the assumption in Proposition 2.6. Any point in the relative interior of the normal cone (in red), such as $\bar{\Lambda}_a$, satisfies the condition, unlike $\bar{\Lambda}_b$ and $\bar{\Lambda}_c$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Note that $g(\Lambda)$ in (8) is indeed the optimal value corresponding to the optimal point in LMO (6). The availability of this oracle motivates us to develop a first-order algorithm for (8). The Lipschitz continuity of $g(\Lambda)$ in (8) is a critical regularity condition to ensure the success of the algorithm. For instance, if $g(\Lambda)$ is Lipschitz continuous and its (sub)gradient is available, then one can use the classical projected gradient ascent with the stepsize proportion to $1/\sqrt{t}$, where t is the iteration count (Shamir & Zhang, 2013). The next result quantifies the Lipschitz constant of $g(\Lambda)$.

Lemma 2.2 (Dual Function Lipschitz Constant). *The function $g(\Lambda)$, defined in (8), is Lipschitz continuous with the constant \mathcal{L} , i.e.,*

$$|g(\Lambda_1) - g(\Lambda_2)| \leq \mathcal{L} \|\Lambda_1 - \Lambda_2\|_{\mathbb{F}} \quad (9)$$

where $\mathcal{L} := \max_{\Sigma \in \mathbb{B}_{\mathbb{S}}^d(\bar{\Sigma})} \|\Sigma\|_{\mathbb{F}}$.

2.2. First-order algorithm and Dykstra projection oracle

We propose a first-order algorithm using the LMO (6) to solve the saddle point problem (7).

Proposition 2.3 (Algorithm & Convergence). *Consider the optimization algorithm*

$$\begin{cases} \Sigma_t = O(\Lambda_t) \\ \Lambda_{t+1} = \Pi_{\mathbb{S}_1 \cap \mathbb{S}_2}[\Lambda_t + \delta \Sigma_t] \\ \bar{\Lambda}_t = \frac{t-1}{t} \bar{\Lambda}_{t-1} + \frac{1}{t} \Lambda_t, \quad \bar{\Sigma}_t = O(\bar{\Lambda}_t) \end{cases} \quad (10)$$

where $\mathbb{S}_1 = \{\Lambda : \text{Diag}(\Lambda) \leq 0\}$, $\mathbb{S}_2 = \{\Lambda : I - \Lambda \in \mathbb{S}_+\}$ are the conic constraint sets of Λ in (7), δ is a constant stepsize, and O is the LMO (6). Starting from any symmetric initial condition Λ_1 , after T iterations we have

$$0 \leq \langle \Lambda^*, \Sigma^* \rangle - \langle \bar{\Lambda}_T, \bar{\Sigma}_T \rangle \leq \frac{\|\Lambda_1 - \Lambda^*\|^2}{2\delta T} + \frac{\delta}{2} \mathcal{L}^2 \quad (11)$$

where (Λ^*, Σ^*) is a saddle point solution of (7), and the constant \mathcal{L} is the Lipschitz constant defined in (9).

Remark 2.4 (Factor Decomposition of $\bar{\Sigma}_T$). The solution $\bar{\Sigma}_T$, obtained by (10), can be decomposed to $L + D$ using the standard factor analysis techniques from Bishop (2006), McLachlan and Krishnan (2007), which is substantially computationally cheaper than its robust counterpart (5). Let us recall that the matrix D gives information about the idiosyncratic noise variances, and the Cholesky factorization of the low-rank matrix L provides the factor loading matrix Φ .

The next remark explains the choice of stepsize in (10).

Remark 2.5 (Diminishing Stepsize & Averaging). Minimizing the error bound in (11) using the constant stepsize δ reveals that the optimal choice is $\delta = \mathcal{O}(1/\sqrt{T})$, yielding a suboptimality gap of $\mathcal{O}(1/\sqrt{T})$; see (Bach, 2014, Corollary 2). However, constant stepsizes often perform poorly in practice. Hence, we adopt a diminishing *anytime* stepsize $\delta_t = \mathcal{O}(1/\sqrt{t})$, which is widely used and achieves similar convergence, up to a logarithmic factor (Shamir & Zhang, 2013).

A possible difficulty in Algorithm (10) is the projection onto $\mathbb{S}_1 \cap \mathbb{S}_2$ for a given symmetric Λ . To address this issue, we leverage Dykstra's algorithm defined as follows:

Algorithm 1. Dykstra's projection ($\Pi_{\mathbb{S}_1 \cap \mathbb{S}_2}[\Lambda]$)

```

1: Input:  $U_2^0 = \Lambda, Z_1^0 = 0, Z_2^0 = 0$ , threshold  $\epsilon$ 
2: while  $\frac{\|U_2^k - U_1^k\|_{\mathbb{F}}}{\|U_2^k\|_{\mathbb{F}}} \geq \epsilon$  do
3:    $U_1^k = \Pi_{\mathbb{S}_1}[U_2^{k-1} + Z_1^{k-1}]$ 
4:    $Z_1^k = U_2^{k-1} + Z_1^{k-1} - U_1^k$ 
5:    $U_2^k = \Pi_{\mathbb{S}_2}[U_1^k + Z_2^{k-1}]$ 
6:    $Z_2^k = U_1^k + Z_2^{k-1} - U_2^k$ 
7:    $k \leftarrow k + 1$ 
return  $U_2^k$ 

```

We know that Algorithm 1 converges *asymptotically* under mild assumptions (Boyle & Dykstra, 1986, Theorem 2). However, this result can be improved to an exponential rate (aka *linear convergence*) for a certain class of conic problems if a corresponding dual admits the Kurdyka-Łojasiewicz property (Wang & Pong, 2024, Theorem 5.3). The next result shows that these conditions are indeed satisfied in our setting.

Proposition 2.6 (Linear Convergence for Projection). *Let $\Lambda^* = \Pi_{\mathbb{S}_1 \cap \mathbb{S}_2}(\bar{\Lambda})$, and assume that $\bar{\Lambda} - \Lambda^*$ resides in the relative interior of the normal cone of $\mathbb{S}_1 \cap \mathbb{S}_2$, i.e., $\bar{\Lambda} - \Lambda^* \in \text{rint}(\mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\Lambda^*))$ (cf. Fig. 1). Then, Algorithm 1 converges linearly to Λ^* from any initial point.*

Remark 2.7 (Relative Interior). Condition $\bar{\Lambda} - \Lambda^* \in \text{rint}(\mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\Lambda^*))$ can be viewed as a genericity condition on $\bar{\Lambda}$. Indeed, we have $\bar{\Lambda} - \Lambda^* \in \mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\Lambda^*)$ for any $\bar{\Lambda}$ (Bauschke & Combettes, 2017, Prop. 6.47). Thus, the assumption excludes that $\bar{\Lambda} - \Lambda^*$ belongs to the boundary, which almost never occurs when $\mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\Lambda^*)$ is not a ray; see Fig. 1. If it is a ray, such as $\bar{\Lambda}_c$ in Fig. 1, then the projection to one of the sets solves the problem, which enables avoiding Dykstra's projection.

3. Special cases of linear minimization oracle

In this section, we study three special cases, the Frobenius distance, the KL divergence, and the Gelbrich distance, of LMO (6), in particular in view of their computational complexity and Lipschitz continuity of (8).

3.1. Frobenius norm

The first distance is the Frobenius norm $F(\Sigma, \hat{\Sigma}) := \|\Sigma - \hat{\Sigma}\|_{\mathbb{F}}$. We show that in this case, the LMO (6) admits an explicit form up to a scalar convex optimization.

Proposition 3.1 (Frobenius Oracle & Lipschitz Constant). *Consider LMO (6) for a given $\hat{\Sigma}$ where the distance function is the Frobenius norm $d(\Sigma, \hat{\Sigma}) = F(\Sigma, \hat{\Sigma})$.*

(i) **Closed-form description:** For any matrix Λ and a positive scalar γ , we define

$$\Sigma^*(\Lambda, \gamma) := \Pi_{\geq 0} \left[\widehat{\Sigma} - \frac{1}{2\gamma} \Lambda \right], \quad (12)$$

where $\Pi_{\geq 0}$ denotes the projection with respect to the Frobenius norm onto the PSD cone. Then, $O(\Lambda)$ in (6) equates to $\Sigma^*(\Lambda, \gamma^*)$, where γ^* is the scalar solution to

$$\max_{0 < \gamma \leq \frac{1}{\varepsilon} \|\Lambda\|_F} \langle \Lambda, \Sigma^*(\Lambda, \gamma) \rangle + \gamma (\|\Sigma^*(\Lambda, \gamma) - \widehat{\Sigma}\|_F^2 - \varepsilon^2). \quad (13)$$

(ii) **Lipschitz constant:** The Lipschitz constant of $g(\Lambda)$ in Lemma 2.2 is bounded by

$$\mathcal{L} \leq \varepsilon + \|\widehat{\Sigma}\|_F \quad (14)$$

Proposition 3.1(i) provides an efficient computational way to implement the LMO (6) since the scalar concave maximization problem (13) can be efficiently solved using bisection within the feasible region $\gamma \in (0, \frac{1}{\varepsilon} \|\Lambda\|_F]$.

3.2. Kullback–Leibler divergence

The second case is the KL divergence between two identical mean normal distributions with covariance matrices Σ and $\widehat{\Sigma}$ (Cover, 1999, Chap. 2), defined as

$$\text{KL}(\Sigma \parallel \widehat{\Sigma}) := \frac{1}{2} \left(-\log \det \Sigma + \log \det \widehat{\Sigma} + \text{Tr}(\Sigma \widehat{\Sigma}^{-1}) - n \right), \quad (15)$$

To make (15) well defined, we assume $\widehat{\Sigma} \succ 0$, and thus invertible. We also assume ξ , α , and ω are Gaussian random vectors. The result can be extended to the case of rank deficient $\widehat{\Sigma}$, where the KL-ball is required to have the same null space. While the KL divergence is a similarity measure, it is not symmetric in its arguments and does not satisfy the triangle inequality. Next, we provide an explicit description of its oracle and dual function Lipschitz constant.

Proposition 3.2 (KL Oracle & Lipschitz Constant). Consider LMO (6) for a given $\widehat{\Sigma}$ where the distance function is the KL divergence $d(\Sigma, \widehat{\Sigma}) = \text{KL}(\Sigma \parallel \widehat{\Sigma})$.

(i) **Closed-form description:** For any matrix Λ and a positive scalar γ , we define

$$\Sigma^*(\Lambda, \gamma) := \left(\widehat{\Sigma}^{-1} + \frac{2}{\gamma} \Lambda \right)^{-1}. \quad (16a)$$

Then, $O(\Lambda)$ in (6) equates to $\Sigma^*(\Lambda, \gamma^*)$, where the scalar γ^* satisfies the equations

$$\begin{cases} \text{KL}(\Sigma^*(\Lambda, \gamma^*) \parallel \widehat{\Sigma}) - \varepsilon = 0 \\ \max \left\{ 0, 2\lambda_{\max}(-\widehat{\Sigma}^{-\frac{1}{2}} \Lambda \widehat{\Sigma}^{-\frac{1}{2}}) \right\} < \gamma^* \leq \\ \|\widehat{\Sigma}^{-\frac{1}{2}} \Lambda \widehat{\Sigma}^{-\frac{1}{2}}\|_* \left(\sqrt{\frac{6}{\varepsilon}} \mathbb{1}_{[0, \frac{1}{24}]} + \left(6 + \frac{1}{4\varepsilon}\right) \mathbb{1}_{(\frac{1}{24}, \infty)} \right) \end{cases} \quad (16b)$$

(ii) **Lipschitz constant:** The Lipschitz constant of $g(\Lambda)$ in Lemma 2.2 is bounded by

$$\mathcal{L} \leq \left(n\sqrt{6\varepsilon} \mathbb{1}_{[0, \frac{1}{24}]} + n\left(6\varepsilon + \frac{1}{4}\right) \mathbb{1}_{(\frac{1}{24}, \infty)} + 1 \right) \|\widehat{\Sigma}\|_F. \quad (17)$$

A similar KL setting as in Proposition 3.2 has been studied in (Ciccone et al., 2018b, Proposition 4.1) where a closed-form solution to the dual of (5) is provided. Proposition 3.2 refines this solution by characterizing the upper and lower bounds (16b) for the optimal dual multiplier. This improvement builds on next lemma that offers a lower bound for $\text{KL}(\Sigma \parallel \widehat{\Sigma})$.

Lemma 3.3 (KL-Lower Bound). The KL divergence $\text{KL}(\Sigma \parallel \widehat{\Sigma})$ in (15) satisfies

$$2 \text{KL}(\Sigma \parallel \widehat{\Sigma}) \geq \sum_{i=1}^n f(\lambda_i(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}})), \quad (18a)$$

where $f(\lambda) := \frac{1}{3}(\lambda-1)^2 \mathbb{1}_{[0, \frac{3}{2}]} + \left(\frac{1}{3}\lambda - \frac{5}{12}\right) \mathbb{1}_{(\frac{3}{2}, \infty)}$. For any radius $\varepsilon \geq 0$, if $\text{KL}(\Sigma \parallel \widehat{\Sigma}) \leq \varepsilon$ we then have

$$|\lambda_{\max}(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) - 1| \leq \sqrt{6\varepsilon} \mathbb{1}_{[0, \frac{1}{24}]} + \left(6\varepsilon + \frac{1}{4}\right) \mathbb{1}_{(\frac{1}{24}, \infty)}. \quad (18b)$$

3.3. Gelbrich distance

The final case is the Gelbrich distance $G(\Sigma, \widehat{\Sigma})$ (Givens & Shortt, 1984), measuring the distance between two distributions with identical means and covariance matrices Σ and $\widehat{\Sigma}$, defined as

$$G(\Sigma, \widehat{\Sigma}) = \sqrt{\text{Tr} \left(\Sigma + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}} \right)}. \quad (19)$$

We adapt the LMO (6) under $G(\Sigma, \widehat{\Sigma})$ from Nguyen et al. (2023), incorporating a slight generalization that extends to arbitrary matrices (rather than PSD matrices as in Nguyen et al., 2023). Additionally, we provide a tight upper bound for the univariate optimization problem and establish the Lipschitz constant of the respective dual function, the two parameters contributing to our algorithm convergence.

Proposition 3.4 (Gelbrich Oracle & Lipschitz Constant). Consider LMO (6) for a given $\widehat{\Sigma} \geq 0$ where the distance function is the Gelbrich distance $d(\Sigma, \widehat{\Sigma}) = G(\Sigma, \widehat{\Sigma})$.

(i) **Closed-form description:** For any Λ and a non-negative scalar γ , we define

$$\Sigma^*(\Lambda, \gamma) := \gamma^2(\gamma I + \Lambda)^{-1} \widehat{\Sigma} (\gamma I + \Lambda)^{-1}. \quad (20a)$$

Then, $O(\Lambda)$ in (6) equates to $\Sigma^*(\Lambda, \gamma^*)$, where γ^* is the unique scalar solution to the concave optimization

$$\begin{cases} \max_{\gamma \geq 0} \gamma \left((I - \gamma(\gamma I + \Lambda)^{-1}, \widehat{\Sigma}) - \varepsilon^2 \right) \\ \text{s.t. } \lambda_{\max}(-\Lambda) < \gamma \leq \frac{1}{\varepsilon} \|\Lambda\|_F (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon) \end{cases} \quad (20b)$$

(ii) **Lipschitz constant:** The Lipschitz constant of $g(\Lambda)$ in Lemma 2.2 is bounded by

$$\mathcal{L} \leq (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon) \varepsilon + \|\widehat{\Sigma}\|_F \quad (21)$$

We note that the quasi-closed form description (20a) and the lower bound for γ in (20b) were previously proposed in (Nguyen et al., 2023, Proposition A.2). However, when Λ is indefinite, which is an important case in the factor model, the upper bound for γ in (20b) and the Lipschitz constant (21) are first introduced here. The proof of Proposition 3.4 builds on next lemma characterizing the regularity properties of the Gelbrich distance and the respective ball.

Lemma 3.5 (Gelbrich Lower Bound). The Gelbrich distance in (19) satisfies

$$G(\Sigma, \widehat{\Sigma}) \geq \max \left\{ \|\lambda^{\frac{1}{2}}(\Sigma) - \lambda^{\frac{1}{2}}(\widehat{\Sigma})\|_2, \lambda_{\max}^{-\frac{1}{2}}(\Sigma + \widehat{\Sigma} + 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \|\Sigma - \widehat{\Sigma}\|_F \right\}. \quad (22a)$$

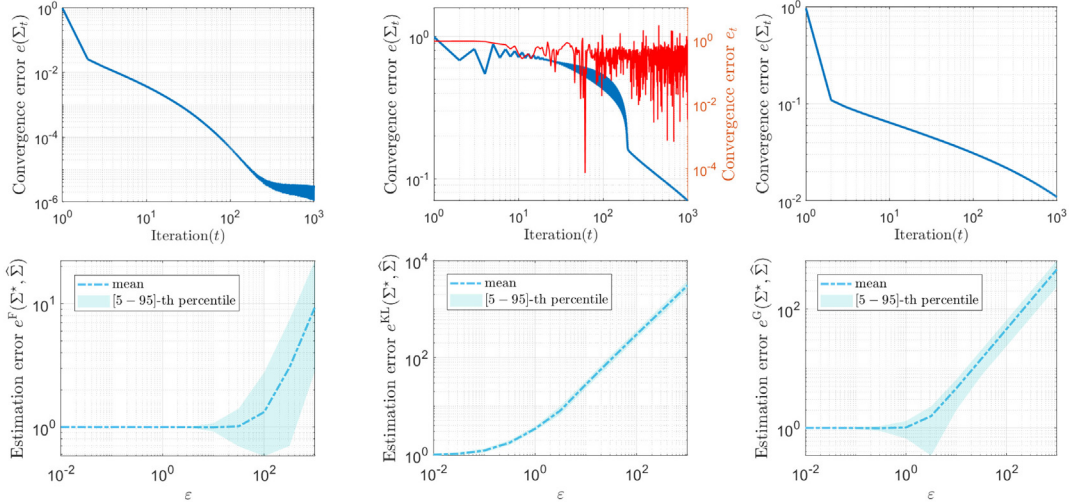


Fig. 2. Analysis of convergence (top row) and estimation of Σ_{True} (bottom row): Frobenius norm (left column), KL divergence (middle column), and Gelbrich distance (right column).

Particularly, for any $\varepsilon \geq 0$, if $G(\Sigma, \widehat{\Sigma}) \leq \varepsilon$, we have

$$G(\Sigma, \widehat{\Sigma}) \geq \frac{\|\Sigma - \widehat{\Sigma}\|_F}{2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon}, \quad (22b)$$

Remark 3.6 (Gelbrich Strong Convexity). The inequality (22b) suggests that the Gelbrich distance is strongly convex with respect to the Frobenius norm, uniformly over any compact subset of the PSD cone, a property of interest in view of optimization algorithms. In comparison with the earlier results in the literature (Bhatia et al., 2018, Theorem 1), the strong convexity relation (22b) does not depend on the minimum eigenvalue of $\widehat{\Sigma}$, making it particularly useful when the ball contains low-rank matrices.

4. Numerical example

In our numerical investigation, we evaluate the performance of our algorithm (10) and observe the effect of ε on covariance matrix estimation accuracy. To this end, we present the numerical results of the algorithm convergence, Σ_{True} estimation, and the algorithm execution time compared to MOSEK. We initialize our algorithm at a random positive definite matrix A_1 projected to $\mathbb{S}_1 \cap \mathbb{S}_2$ using Dykstra’s projection. The algorithm stopping condition is the normalized relative change as

$$\frac{|\langle A_t, \Sigma_t \rangle - \langle A_{t-1}, \Sigma_{t-1} \rangle|}{|\langle A_t, \Sigma_t \rangle|} \leq 10^{-6}. \quad (23)$$

4.1. Synthetic data generation

We generate the data in two steps: First, we build the ground-truth covariance matrix $\Sigma_{\text{True}} = D_{\text{True}} + L_{\text{True}}$, where $L_{\text{True}} = \Phi_{\text{True}} \Phi_{\text{True}}^T$, and Φ_{True} and D_{True} are created pseudorandomly using *rand* function in MATLAB with *rng(1, ‘twister’)* and *rng(0)*, respectively. We note that Φ_{True} and D_{True} are adjusted to be far away from the origin, by setting a lower bound for their minimum element (we set the bound to be 5, by adding 5 to *rand* function) so that Σ_{True} is far from 0, to exclude the trivial solution $\Sigma = 0$, when $\Lambda \geq 0$ and the ball includes 0. In the next step, we build $N = 15n$ pseudorandom samples of factors $\{\alpha_k\}_{k=1}^N$ and noise $\{\omega_k\}_{k=1}^N$, from zero mean normal distributions with covariance matrices $\Sigma_\alpha = I$ and D_{True} , respectively. Then, we build the observed vectors samples $\{\xi_k = \Phi_{\text{True}}\alpha_k + w_k\}_{k=1}^N$ and compute $\widehat{\Sigma}$.

4.2. Convergence

The saddle point problem (7) is solved using (10) with $\delta_t = 1/\sqrt{t}$ stepsize, 10^4 iterations, for the heart disease dataset at www.kaggle.com/datasets/johnsmith88/heart-disease-dataset, downloaded on April 2026, with $n = 13$ including age, chest pain type, and resting blood pressure to indicate the presence of heart disease, for $F(\Sigma, \widehat{\Sigma})$, $KL(\Sigma \parallel \widehat{\Sigma})$, and $G(\Sigma, \widehat{\Sigma})$ cases, and for $\varepsilon = \sqrt{10}$, $\varepsilon = 0.01$, and $\varepsilon = 0.1$, respectively. Since the true optimal value is unknown, we choose the objective value at iteration 10^4 as the reference optimal objective value $\langle \Lambda^*, \Sigma^* \rangle$ and investigate the algorithm convergence by the normalized convergence error as

$$e(\Sigma_t) := \frac{|\langle \Lambda_t, \Sigma_t \rangle - \langle \Lambda^*, \Sigma^* \rangle|}{|\langle \Lambda^*, \Sigma^* \rangle|}, \quad (24)$$

We note that to prevent terminating earlier than 10^4 iterations, stopping condition (23) is not activated. Fig. 2, top row demonstrates that after 200 iterations, the normalized error (24) decreases to around 7.8×10^{-6} , 0.16, and 0.02 of its initial value, for $F(\Sigma, \widehat{\Sigma})$, $KL(\Sigma \parallel \widehat{\Sigma})$, and $G(\Sigma, \widehat{\Sigma})$ cases, respectively. Hence, the theoretical convergence result of Proposition 2.3 is validated. Moreover, based on the red curve in the KL case, which shows convergence of Ciccone et al. (2018b) with $\varepsilon = 0.01$, our algorithm outperforms the alternating direction method of multipliers (ADMM) algorithm employed in Ciccone et al. (2018b). In this case, the normalized convergence error is defined as $e_t := \frac{|F(\lambda_t, X_t) - F(\lambda^*, X^*)|}{|F(\lambda^*, X^*)|}$, with F , λ , and X as in Ciccone et al. (2018b).

4.3. Estimation of the ground-truth Σ_{True}

The effect of the hyperparameter ε on Σ_{True} estimation error is investigated for $N_{\text{exp}} = 100$ experiments, for $n = 20$ and $r = 4$. In each experiment, the factor model problem is solved using Algorithm (10) with $\delta_t = 1/\sqrt{t}$ stepsize for $\varepsilon \in \Omega = \{0.01(\sqrt{10})^i \mid i = 0, 1, \dots, 10\}$ and for maximum $t_{\text{end}} = 10^4$ iterations, while the stopping criterion (23) is active. The data is generated through the synthetic data generation procedure in Section 4.1 with fixed D_{True} and Σ_{True} for all experiments. In each experiment, the same samples of observed vectors $\{\xi_k\}_{k=1}^N$ are used to solve the problem for various values of ε , while these samples vary among different experiments. We define the normalized estimation error of Σ_{True} as

$$e^d(\Sigma^*, \widehat{\Sigma}) := \frac{d(\Sigma^*, \Sigma_{\text{True}})}{d(\widehat{\Sigma}, \Sigma_{\text{True}})}. \quad (25)$$

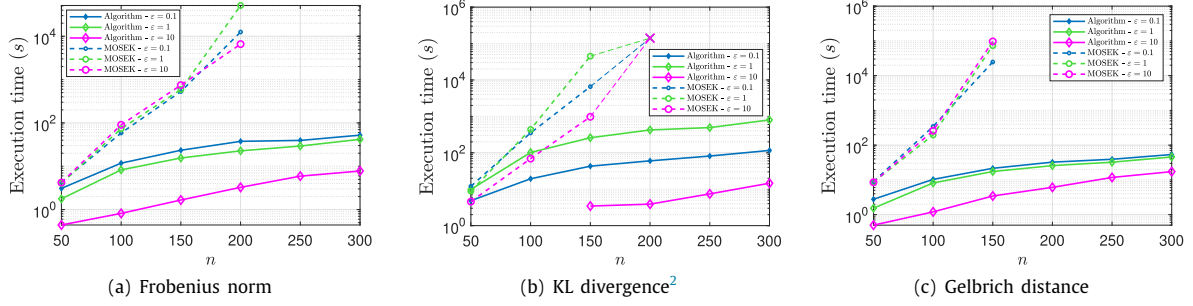


Fig. 3. Computational time comparison of algorithm (10) and MOSEK.

For $F(\Sigma, \widehat{\Sigma})$ and $G(\Sigma, \widehat{\Sigma})$ cases, an improvement in Σ_{True} estimation compared to $\widehat{\Sigma}$ is achieved in 61% and 52% of the experiments, respectively, as shown as sweet spots in Fig. 2, bottom row, for $\varepsilon = 100$ and $\varepsilon = \sqrt{10}$. For $\text{KL}(\Sigma \parallel \widehat{\Sigma})$ case, while a sweet spot is not seen, Σ_{True} estimation is slightly improved in 37% of the experiments.

4.4. Execution time

For $F(\Sigma, \widehat{\Sigma})$, $\text{KL}(\Sigma \parallel \widehat{\Sigma})$, and $G(\Sigma, \widehat{\Sigma})$ cases, the execution time of our algorithm with stopping condition (23), maximum 10^6 iterations, and $\delta_t = 1/\sqrt{t}$ is compared to MOSEK, for $\varepsilon \in \{0.1, 1, 10\}$ and synthetic datasets generated as in 4.1. We solve (5) and reformulations in Lemmas B.1 and B.2 for $F(\Sigma, \widehat{\Sigma})$, $\text{KL}(\Sigma \parallel \widehat{\Sigma})$, and $G(\Sigma, \widehat{\Sigma})$, respectively, for implementations using MOSEK 9.3 with YALMIP (Lofberg, 2004) in MATLAB R2024a, on a Core(TM) i7-1185G7 CPU with 3.00 GHz (1.80 GHz) clock speed and 16 GB RAM. According to Fig. 3, our algorithm is clearly more efficient than MOSEK, especially for higher-dimensional data. MOSEK failed to solve the $F(\Sigma, \widehat{\Sigma})$ and $\text{KL}(\Sigma \parallel \widehat{\Sigma})$ cases with $n \geq 250$ and the $G(\Sigma, \widehat{\Sigma})$ case with $n \geq 200$, due to running out of memory. We note that for $F(\Sigma, \widehat{\Sigma})$ and $G(\Sigma, \widehat{\Sigma})$ cases, the execution time for bigger ε is less than smaller ε . For $\text{KL}(\Sigma \parallel \widehat{\Sigma})$ case, the effect of increasing ε on the execution time is not clear.

5. Conclusion and future direction

We propose a saddle point reformulation of the factor model problem and a first-order algorithm relying on an LMO. We address the inaccuracy of the empirical covariance matrix $\widehat{\Sigma}$ by robustification via a family of covariance matrices around it with respect to a generic distance function. We further derive semi-closed form solution for the LMO in three distance functions: Frobenius norm, KL divergence, and Gelbrich distance. For future research, the physical meaning of the factors and factor loading matrix in dynamical systems can be found, which helps to predict system behavior by analyzing these components, or by identifying their variations to investigate stability or detect potential faults. Controllers can then be designed by finding the mapping between control parameters and factor model components.

² For our algorithm, for $\varepsilon = 10$, $n = 50$ and $n = 100$, since the optimal objective value is 0, the stopping criterion (23) is never triggered. Therefore, for these cases we set the iteration limit to 10^4 , yielding optimal objective values of -3.507 and -0.0087 , respectively. These cases were executed in another, similar, hardware. For $n = 200$, MOSEK could not converge after 39 h, for all three ε values.

Appendix A. Technical proofs

Proof of Proposition 2.1. Let us consider a new decision variable $\Sigma = L + D$, aligned with (2). The corresponding Lagrange multiplier of this equality constraint is denoted by Λ , which is symmetric because Σ , L , and D are symmetric. The factor model problem (5) can then be rewritten as

$$\begin{aligned} \max_{\Lambda} \quad & \min_{\Sigma, L, D} \quad \text{Tr}(L) + \langle \Lambda, \Sigma - L - D \rangle \\ \text{s.t.} \quad & L \in \mathbb{S}_+, \quad D \in \mathbb{D}_+, \quad \Sigma \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma}), \end{aligned} \quad (\text{A.1})$$

where the strong duality holds thanks to the usual convex-concave property. We note that the Slater's condition is not required due to the linearity of the dualized constraint (Bertsekas, 2009). The decision variables L and D in (A.1) are separable, and as such, the inner minimization over their respective conic spaces \mathbb{S}_+ and \mathbb{D}_+ can be computed explicitly. This yields the so-called support functions $\max_{L \in \mathbb{S}_+} \langle \Lambda - I, L \rangle$, and $\max_{D \in \mathbb{D}_+} \langle \Lambda, D \rangle$ in the objective. It is worth stating that using the dual cone definition, $\max_{L \in \mathbb{S}_+} \langle \Lambda - I, L \rangle = 0$, if $I - \Lambda \in \mathbb{S}_+^*$; otherwise is $+\infty$; and $\max_{D \in \mathbb{D}_+} \langle \Lambda, D \rangle = 0$ if $-\Lambda \in \mathbb{D}_+^*$; otherwise is $+\infty$; for more general setting, the reader is referred to (Bertsekas, 2009, Proposition 5.3.9). Hence, the support functions essentially confine the feasible set of $I - \Lambda$ and $-\Lambda$ to the respective dual cones \mathbb{S}_+^* and \mathbb{D}_+^* , respectively. It then suffices to note that the PSD cone is self-dual (i.e., $\mathbb{S}_+^* = \mathbb{S}_+$) to arrive at the desired program (7). \square

Proof of Lemma 2.2. Using the definition of the dual function (8), we have

$$\begin{aligned} g(\Lambda_1) - g(\Lambda_2) &= \min_{\Sigma_1 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda_1, \Sigma_1 \rangle - \min_{\Sigma_2 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda_2, \Sigma_2 \rangle \\ &= \min_{\Sigma_1 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \max_{\Sigma_2 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda_1, \Sigma_1 \rangle - \langle \Lambda_2, \Sigma_2 \rangle \\ &\leq \max_{\Sigma_2 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda_1 - \Lambda_2, \Sigma_2 \rangle \leq \mathcal{L} \|\Lambda_1 - \Lambda_2\|_F, \end{aligned}$$

where the second equality is derived since for every function $f(x)$, we have $-\min_x f(x) = \max_x -f(x)$, and since $\langle \Lambda_1, \Sigma_1 \rangle$ is constant when maximizing over Σ_2 . The first inequality holds since the minimum value of the function $\max_{\Sigma_2 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda_1, \Sigma_1 \rangle - \langle \Lambda_2, \Sigma_2 \rangle$, over Σ_1 , is less than or equal to its value at any other feasible point including Σ_2 . The last inequality is the direct application of the Cauchy-Schwarz inequality (i.e., $\langle A, B \rangle \leq \|A\|_F \|B\|_F$ for all A, B with appropriate dimensions). \square

Proof of Proposition 2.3. The proof follows from the standard projected subgradient method. LMO (6) is the subgradient of dual function (8), as a consequence of Danskin's theorem (Bertsekas, 1971, (A.22), p 154), i.e., $O(\Lambda) = \frac{\partial g(\Lambda)}{\partial \Lambda}$. Hence, the first two steps in (10) effectively implement the projected subgradient ascent

applied to (8) while $\bar{\Lambda}_t$ is the average over all the iterations, i.e., $\bar{\Lambda}_t = \frac{1}{N} \sum_{i=1}^t \Lambda_i$. Leveraging the classical bound of the projected subgradient (e.g., (Bertsekas, 2015, Proposition 3.2.4)), we have the error bound (11). \square

Proof of Proposition 2.6. It suffices to show all requirements of (Wang & Pong, 2024, Theorem 5.3) specified in (Wang & Pong, 2024, Assumption 4.1) are satisfied for $\Pi_{\mathbb{S}_1 \cap \mathbb{S}_2}(\Lambda)$. Since \mathbb{S}_1 is a polyhedron, it is C^2 -cone reducible in the sense of (Bonnans & Shapiro, 2013, Definition 3.135). Similarly, \mathbb{S}_2 can be reduced to a (shifted) semidefinite cone. Thus, the first requirement is satisfied. The second and third requirements also hold as $\text{rint}(\mathbb{S}_1 \cap \mathbb{S}_2)$ is nonempty and due to our assumption that $\Lambda - \Lambda^* \in \text{rint}(\mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\Lambda^*))$. \square

Proof of Proposition 3.1. Regarding part (i), by introducing the Lagrange multiplier γ for the constraint $\|\Sigma - \widehat{\Sigma}\|_F^2 \leq \varepsilon^2$, the dual program of the oracle (6) is

$$\max_{\gamma \geq 0} \min_{\Sigma \geq 0} \langle \Lambda, \Sigma \rangle + \gamma (\|\Sigma - \widehat{\Sigma}\|_F^2 - \varepsilon^2). \quad (\text{A.2})$$

Since the Hessian of the Lagrangian function with respect to Σ is a scaled identity, we first compute the solution of the unconstrained program of the inner minimization (ignoring $\Sigma \geq 0$), and then project it onto the PSD cone. The first-order optimality condition for the unconstrained inner minimization yields $\Lambda + 2\gamma(\Sigma^* - \widehat{\Sigma}) = 0$, for every (Λ, γ) , which describes the unique solution to this problem, yielding the closed-form projected solution $\Sigma^*(\Lambda, \gamma)$ in (12). Substituting $\Sigma^*(\Lambda, \gamma)$ in (A.2) arrives at (13) whose solution γ^* together with the closed-form solution $\Sigma^*(\Lambda, \gamma^*)$ yields the saddle point of (A.2). We note that excluding $\gamma = 0$ is without loss of generality since the objective function of the inner minimization is lower-semicontinuous in γ (pointwise minimum of a continuous function in both variables (γ, Λ)). To show that γ^* is bounded by $\|\Lambda\|_F$, we use (Nedić & Ozdaglar, 2009, Lemma 1), which relies on the existence of a Slater point. Considering $\widehat{\Sigma}$ as the Slater point, (Nedić & Ozdaglar, 2009, Lemma 1) offers the upper bound

$$\begin{aligned} \gamma^* &\leq \frac{1}{\varepsilon^2} \left(\langle \widehat{\Sigma}, \Lambda \rangle - \min_{\Sigma \in \mathbb{B}_F^{\varepsilon}(\widehat{\Sigma})} \langle \Sigma, \Lambda \rangle \right) \\ &= \max_{\Sigma \in \mathbb{B}_F^{\varepsilon}(\widehat{\Sigma})} \frac{1}{\varepsilon^2} \langle \widehat{\Sigma} - \Sigma, \Lambda \rangle \leq \frac{1}{\varepsilon} \|\Lambda\|_F \end{aligned} \quad (\text{A.3})$$

where the last inequality is a consequence of lifting the constraint $\Sigma \geq 0$ and the fact that the Frobenius norm is self-dual. Next, we prove part (ii). Thanks to Lemma 2.2, in particular (9), the Lipschitz constant \mathcal{L} satisfies

$$\mathcal{L} \leq \max_{\Sigma \in \mathbb{B}_F^{\varepsilon}(\widehat{\Sigma})} \|\Sigma - \widehat{\Sigma}\|_F + \|\widehat{\Sigma}\|_F \leq \varepsilon + \|\widehat{\Sigma}\|_F$$

where the first inequality is the basic triangle inequality of the norm, and the second inequality is the direct consequence of the constraint $\Sigma \in \mathbb{B}_F^{\varepsilon}(\widehat{\Sigma})$. \square

Proof of Proposition 3.2. The proof of both parts follows similar lines as Proposition 3.1. Concerning (i), dualizing the distance function constraint corresponding to $\text{KL}(\Sigma \parallel \widehat{\Sigma}) \leq \varepsilon$ in the LMO (6) yields

$$\begin{aligned} \max_{\gamma \geq 0} \min_{\Sigma \geq 0} \langle \Lambda, \Sigma \rangle + \frac{\gamma}{2} \left(-\log \det \Sigma + \log \det \widehat{\Sigma} \right. \\ \left. + \text{Tr}(\Sigma \widehat{\Sigma}^{-1}) - n \right) - \gamma \varepsilon. \end{aligned} \quad (\text{A.4})$$

Due to the term $-\log \det \Sigma$ in the objective of (A.4), we can exclude the boundary of the PSD cone for the inner optimizer Σ (i.e., $\Sigma \succ 0$), when $\gamma > 0$. Similar to proof of Proposition

3.1, we can exclude $\gamma = 0$ since the optimal value of the inner minimization is lower-semicontinuous in γ . Thus, the first-order optimality condition for the inner minimization yields the unique minimizer $\Sigma^*(\Lambda, \gamma)$ in (16a), substituting which in (A.4) yields

$$\begin{aligned} \max_{\gamma > 0} \langle \Lambda, \Sigma^*(\Lambda, \gamma) \rangle + \frac{\gamma}{2} \left(-\log \det \Sigma^*(\Lambda, \gamma) \right. \\ \left. + \log \det \widehat{\Sigma} + \text{Tr}(\Sigma^*(\Lambda, \gamma) \widehat{\Sigma}^{-1}) - n \right) - \gamma \varepsilon. \end{aligned} \quad (\text{A.5})$$

Applying the first-order optimality condition to (A.5) yields the algebraic equation (16b). Next, we derive the lower bound for γ^* solving (A.5) (or equivalently (16b)). Since $\Sigma^*(\Lambda, \gamma^*) \succ 0$, we have $\widehat{\Sigma}^{-1} + \frac{2}{\gamma^*} \Lambda \succ 0$, which implies $\gamma^* > 2\lambda_{\max}(-\widehat{\Sigma}^{-\frac{1}{2}} \Lambda \widehat{\Sigma}^{-\frac{1}{2}})$. This bound and the original non-negativity constraint conclude the lower bound. To show the upper bound of γ^* in (16b), following similar lines as in Proposition 3.1, we have

$$\begin{aligned} \gamma^* &\leq \max_{\Sigma \in \mathbb{B}_{\text{KL}}^{\varepsilon}(\widehat{\Sigma})} \frac{1}{\varepsilon} \langle \widehat{\Sigma} - \Sigma, \Lambda \rangle \leq \begin{cases} \max_{\Sigma \geq 0} \frac{1}{\varepsilon} \langle \widehat{\Sigma} - \Sigma, \Lambda \rangle \\ \text{s.t. } \text{KL}(\Sigma \parallel \widehat{\Sigma}) \leq \varepsilon \end{cases} \\ &\leq \begin{cases} \max_{\Sigma \geq 0} \frac{1}{\varepsilon} \langle \widehat{\Sigma} - \Sigma, \Lambda \rangle \\ \text{s.t. } \left| \lambda_{\max}(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) - 1 \right| \\ \leq \sqrt{6\varepsilon} \mathbb{1}_{[0, \frac{1}{24}]} + (6\varepsilon + \frac{1}{4}) \mathbb{1}_{(\frac{1}{24}, \infty)}, \end{cases} \end{aligned} \quad (\text{A.6})$$

where the last inequality follows from (18b). The objective function in (A.6) can be upper bounded by

$$\begin{aligned} \langle \widehat{\Sigma} - \Sigma, \Lambda \rangle &= \langle I - \widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}, \widehat{\Sigma}^{\frac{1}{2}} \Lambda \widehat{\Sigma}^{\frac{1}{2}} \rangle \\ &\leq \max_{i \leq n} |\lambda_i(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) - 1| \|\widehat{\Sigma}^{\frac{1}{2}} \Lambda \widehat{\Sigma}^{\frac{1}{2}}\|_*, \end{aligned}$$

Considering the upper bound in (A.6) and replacing λ_{\max} yields the desired upper bound for γ^* in (16b). Regarding \mathcal{L} in part (ii), following bound (9), we have

$$\begin{aligned} \mathcal{L} &= \max_{\Sigma \in \mathbb{B}_{\text{KL}}^{\varepsilon}(\widehat{\Sigma})} \|\Sigma\|_F \leq \max_{\Sigma \in \mathbb{B}_{\text{KL}}^{\varepsilon}(\widehat{\Sigma})} \|\Sigma - \widehat{\Sigma}\|_F + \|\widehat{\Sigma}\|_F \\ &\leq \left(\max_{\Sigma \in \mathbb{B}_{\text{KL}}^{\varepsilon}(\widehat{\Sigma})} \|\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}} - I\|_F + 1 \right) \|\widehat{\Sigma}\|_F \\ &\leq \left(\max_{\Sigma \in \mathbb{B}_{\text{KL}}^{\varepsilon}(\widehat{\Sigma})} n \lambda_{\max}(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}} - I) + 1 \right) \|\widehat{\Sigma}\|_F \\ &\leq \left(n \sqrt{6\varepsilon} \mathbb{1}_{[0, \frac{1}{24}]} + n(6\varepsilon + \frac{1}{4}) \mathbb{1}_{(\frac{1}{24}, \infty)} + 1 \right) \|\widehat{\Sigma}\|_F, \end{aligned}$$

where the last inequality follows from (18b). \square

Proof of Lemma 3.3. To prove (18a), following (15), we have

$$\begin{aligned} 2 \text{KL}(\Sigma \parallel \widehat{\Sigma}) &= \text{Tr} \left(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}} - I \right) - \log \det(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) \\ &= \sum_{i=1}^n \lambda_i(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) - 1 - \log \lambda_i(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) \\ &\geq \sum_{i=1}^n f(\lambda_i(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}})), \end{aligned}$$

where the equalities hold, since the trace and determinant operators are symmetric (i.e., $\text{Tr}(AB) = \text{Tr}(BA)$ and $\det(AB) = \det(BA)$). The inequality follows from the Taylor series expansion with degree 2 of the convex function $\lambda - 1 - \log(\lambda)$ at $\lambda = 1$ within the interval $[0, \frac{3}{2}]$, followed by a linear extension of this lower bound for $\lambda > \frac{3}{2}$. To prove (18b), we use an inverse function argument for the lower bound function f in (18a) and apply the inequality $\max_{i \leq n} f(\lambda_i) \leq \sum_{i \leq n} f(\lambda_i)$. \square

Proof of Proposition 3.4. Regarding (20a) and the lower bound of γ in (20b), we refer to parts (i) and (ii) in (Nguyen et al., 2023,

Proposition A.2). Concerning the upper bound in (20b), we follow the approach used for $F(\Sigma, \widehat{\Sigma})$ to derive a similar bound to (A.3) as

$$\begin{aligned} \gamma^* &\leq \max_{\Sigma \in \mathbb{B}_\varepsilon^G(\widehat{\Sigma})} \frac{1}{\varepsilon^2} (\widehat{\Sigma} - \Sigma, \Lambda) = \begin{cases} \max_{\Sigma \geq 0} & \frac{1}{\varepsilon^2} (\widehat{\Sigma} - \Sigma, \Lambda) \\ \text{s.t.} & G(\Sigma, \widehat{\Sigma}) \leq \varepsilon \end{cases} \\ &\leq \begin{cases} \max_{\Sigma \geq 0} & \frac{1}{\varepsilon^2} \|\Lambda\|_F \|\widehat{\Sigma} - \Sigma\|_F \\ \text{s.t.} & \|\Sigma - \widehat{\Sigma}\|_F (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon)^{-1} \leq \varepsilon \end{cases} \\ &\leq \frac{1}{\varepsilon} \|\Lambda\|_F (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon), \end{aligned}$$

where the second inequality is the application of Cauchy–Schwarz inequality to the objective function and (22b) to $G(\Sigma, \widehat{\Sigma})$ in the constraint. Regarding \mathcal{L} in (21), we apply (22b) to the Lipschitz constant (9) to arrive at

$$\begin{aligned} \mathcal{L} &= \max_{\Sigma \in \mathbb{B}_\varepsilon^G(\widehat{\Sigma})} \|\Sigma\|_F \leq \max_{\Sigma \in \mathbb{B}_\varepsilon^G(\widehat{\Sigma})} \|\Sigma - \widehat{\Sigma}\|_F + \|\widehat{\Sigma}\|_F \\ &\leq (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon)\varepsilon + \|\widehat{\Sigma}\|_F. \quad \square \end{aligned}$$

Proof of Lemma 3.5. Regarding the first term on the right-hand side of inequality (22a), note that

$$\begin{aligned} G^2(\Sigma, \widehat{\Sigma}) &= \text{Tr} \left(\Sigma + \widehat{\Sigma} - 2 \left(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \\ &= \sum_{i \leq n} \lambda_i(\Sigma) + \lambda_i(\widehat{\Sigma}) - 2\lambda_i^{\frac{1}{2}}(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}) \\ &\geq \sum_{i \leq n} \lambda_i(\Sigma) + \lambda_i(\widehat{\Sigma}) - 2\lambda_i^{\frac{1}{2}}(\widehat{\Sigma})\lambda_i^{\frac{1}{2}}(\Sigma) \quad (\text{A.7}) \\ &= \sum_{i \leq n} (\lambda_i^{\frac{1}{2}}(\Sigma) - \lambda_i^{\frac{1}{2}}(\widehat{\Sigma}))^2 = \|\lambda^{\frac{1}{2}}(\Sigma) - \lambda^{\frac{1}{2}}(\widehat{\Sigma})\|_2^2, \end{aligned}$$

where the inequality follows from the matrix version of Hardy–Littlewood–Polya inequality (Carlen & Lieb, 2006, Theorem 3.2). As for the second term on the right-hand side of (22a),

$$\begin{aligned} &\text{Tr} \left((\Sigma + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) (\Sigma + \widehat{\Sigma} + 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \right) \\ &= \text{Tr} \left((\Sigma + \widehat{\Sigma})^2 - 4(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}) \right) \\ &= \text{Tr} (\Sigma^2 - 2\Sigma\widehat{\Sigma} + \widehat{\Sigma}^2) = \text{Tr} ((\Sigma - \widehat{\Sigma})^2) = \|\Sigma - \widehat{\Sigma}\|_F^2, \end{aligned}$$

where the symmetric property $\text{Tr}(AB) = \text{Tr}(BA)$ is used to derive the first and second equalities. Using Von Neumann’s trace inequality on the above equality gives $G^2(\Sigma, \widehat{\Sigma})\lambda_{\max}(\widehat{\Sigma} + \Sigma + 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \geq \|\Sigma - \widehat{\Sigma}\|_F^2$, which, together with (A.7) concludes (22a). Now observe $\varepsilon \geq G(\Sigma, \widehat{\Sigma}) \geq \|\lambda^{\frac{1}{2}}(\Sigma) - \lambda^{\frac{1}{2}}(\widehat{\Sigma})\|_2 \geq |\lambda_{\max}^{\frac{1}{2}}(\Sigma) - \lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma})|$ which implies $\lambda_{\max}^{\frac{1}{2}}(\Sigma) \leq \lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon$. Therefore

$$\begin{aligned} &\lambda_{\max}(\Sigma + \widehat{\Sigma} + 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &\leq \lambda_{\max}(\Sigma + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) + \lambda_{\max}(4(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &\leq G^2(\Sigma, \widehat{\Sigma}) + 4\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma})\lambda_{\max}^{\frac{1}{2}}(\Sigma) \\ &\leq \varepsilon^2 + 4\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma})(\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon) = (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon)^2. \end{aligned}$$

The first two inequalities hold due to sub-additivity and sub-multiplicativity of λ_{\max} , together with the property $\lambda_{\max}(A) \leq \text{Tr}(A)$ in the PSD cone. We apply the above bound in (22a) to arrive at (22b). \square

Appendix B. Technical lemmas to solve (5) using MOSEK

Lemma B.1 (KL-Factor Model Reformulation). The factor model problem (5) with the KL divergence $d(\Sigma, \widehat{\Sigma}) = \text{KL}(\Sigma \parallel \widehat{\Sigma})$ is equivalent to the program

$$\begin{aligned} &\min_{L, \Sigma, Z} \langle L, I \rangle \\ &\text{s.t.} \quad \begin{bmatrix} \Sigma & Z \\ Z^\top & \text{Diag}(Z) \end{bmatrix} \succeq 0, \quad L \in \mathbb{S}_+, \quad Z \in \mathbb{L}_n, \quad \Sigma - L \in \mathbb{D}_+ \\ &\quad \log(\det \widehat{\Sigma}) + \text{Tr}(\Sigma \widehat{\Sigma}^{-1}) - n - 2\varepsilon \leq \sum_i \log(Z_{ii}). \end{aligned}$$

For the proof, we refer to (Mosek ApS, 2025, Section 6.2.3).

Lemma B.2 (Gelbrich-Factor Model Reformulation). The factor model problem (5) with $d(\Sigma, \widehat{\Sigma}) = G(\Sigma, \widehat{\Sigma})$ is equivalent to the LMI program

$$\begin{aligned} &\min_{L, \Sigma, C} \langle L, I \rangle \\ &\text{s.t.} \quad \begin{bmatrix} \Sigma & C \\ C^\top & \widehat{\Sigma} \end{bmatrix} \succeq 0, \quad \Sigma - L \in \mathbb{D}_+, \quad L \in \mathbb{S}_+ \\ &\quad \langle \Sigma + \widehat{\Sigma} - 2C, I \rangle \leq \varepsilon^2 \end{aligned}$$

The proof is an application of (Nguyen et al., 2023, Proposition 2.2) to (5).

References

- Anderson, Theodore Wilbur (2003). *An introduction to multivariate statistical analysis*. Wiley.
- Bach, Francis (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1), 595–627.
- Bauschke, Heinz H., & Combettes, Patrick L. (2017). *Convex analysis and monotone operator theory in Hilbert spaces* (2nd ed.) Springer.
- Bertsekas, Dimitri (1971). *Control of uncertain systems with a set-membership description of the uncertainty* (Ph.D. thesis), Massachusetts Institute of Technology.
- Bertsekas, Dimitri (2009). *Convex optimization theory*. Athena Scientific.
- Bertsekas, Dimitri (2015). *Convex optimization algorithms*. Athena Scientific.
- Bertsimas, Dimitris, Copenhaver, Martin S., & Mazumder, Rahul (2017). Certifiably optimal low rank factor analysis. *Journal of Machine Learning Research*, 18, 1–53.
- Bhatia, Rajendra, Jain, Tanvi, & Lim, Yongdo (2018). Strong convexity of sandwiched entropies and related optimization problems. *Reviews in Mathematical Physics*, 30(09), Article 1850014.
- Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Springer.
- Bonnans, Joseph Frédéric, & Shapiro, Alexander (2013). *Perturbation analysis of optimization problems*. Springer.
- Bottegal, Giulio, & Picci, Giorgio (2014). Modeling complex systems by generalized factor analysis. *TACON*, 60(3), 759–774.
- Boyle, James P., & Dykstra, Richard L. (1986). A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in order restricted statistical inference* (pp. 28–47). Springer.
- Carlen, Eric, & Lieb, Elliott H. (2006). Some matrix rearrangement inequalities. *Annali Di Matematica Pura Ed Applicata*, 185(Suppl 5), S315–S324.
- Ciccione, Valentina, Ferrante, Augusto, & Zorzi, Mattia (2018a). An alternating minimization algorithm for factor analysis. arXiv:1806.04433.
- Ciccione, Valentina, Ferrante, Augusto, & Zorzi, Mattia (2018b). Factor models with real data: A robust estimation of the number of factors. *TACON*, 64(6), 2412–2425.
- Cover, Thomas M. (1999). *Elements of information theory*. Wiley.
- Donoho, David L. (2006). Compressed sensing. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 52(4), 1289–1306.
- Engle, Robert, & Watson, Mark (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76(376), 774–781.
- Falconi, Lucia, Ferrante, Augusto, & Zorzi, Mattia (2023). A robust approach to ARMA factor modeling. *IEEE Transactions on Automatic Control*, 69(2), 828–841.
- Forni, Mario, & Lippi, Marco (2001). The generalized dynamic factor model: Representation theory. *Econometric Theory*, 17(6), 1113–1141.
- Givens, Clark R., & Shortt, Rae Michael (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2), 231–240.
- Khamaru, Koulik, & Mazumder, Rahul (2019). Computation of the maximum likelihood estimator in low-rank factor analysis. *Mathematical Programming*, 176, 279–310.

- Lofberg, Johan (2004). YALMIP: A toolbox for modeling and optimization in MATLAB. In *ICRA 2004* (pp. 284–289).
- McLachlan, Geoffrey J., & Krishnan, Thriyambakam (2007). *The EM algorithm and extensions*. Wiley.
- Mosek ApS (2025). MOSEK modeling cookbook: Release 3.3.1.
- Nedić, Angelia, & Ozdaglar, Asuman (2009). Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4), 1757–1780.
- Nguyen, Viet Anh, Shafieezadeh-Abadeh, Soroosh, Kuhn, Daniel, & Mohajerin Esfahani, Peyman (2023). Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 48(1), 1–37.
- O'Donoghue, Brendan, Chu, Eric, Parikh, Neal, & Boyd, Stephen (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3), 1042–1068.
- Picci, Giorgio, Falconi, Lucia, Ferrante, Augusto, & Zorzi, Mattia (2023). Hidden factor estimation in dynamic generalized factor analysis models. *Automatica*, 149, Article 110834.
- Recht, Benjamin, Fazel, Maryam, & Parrilo, Pablo A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3), 471–501.
- van Schuppen, Jan Hendrik (1986). Stochastic realization problems motivated by econometric modeling. In *Modelling, identification and robust control* (pp. 259–275).
- Shamir, Ohad, & Zhang, Tong (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning* (pp. 71–79). PMLR.
- Spearman, Charles (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292.
- Wang, Xiaozhou, & Pong, Ting Kei (2024). Convergence rate analysis of a Dykstra-type projection algorithm. *SIAM Journal on Optimization*, 34(1), 563–589.
- Wu, Ningning, & Zhang, Jing (2006). Factor-analysis based anomaly detection and clustering. *Decision Support Systems*, 42(1), 375–389.
- Zhou, Le, Wang, Yaixin, Ge, Zhiqiang, & Song, Zhihuan (2018). Multirate factor analysis models for fault detection in multirate processes. *IEEE Transactions on Industrial Informatics*, 15(7), 4076–4085.
- Zorzi, Mattia, & Sepulchre, Rodolphe (2015). Factor analysis of moving average processes. In *ECC* (pp. 3579–3584).



Shabnam Khodakaramzadeh has received her B.Sc. and M.Sc. degrees in Mechanical Engineering from Sharif University of Technology, Tehran, Iran and University of Tehran, Tehran, Iran in 2016 and 2018, respectively.

She is a Ph.D. candidate at Delft University of Technology, Delft, The Netherlands. Her research area includes optimization, control, and fault detection and estimation.



Soroosh Shafiee is an assistant professor in the School of Operations Research and Information Engineering at Cornell University. Before that, he held positions as a postdoctoral researcher at both the Tepper School of Business at Carnegie Mellon University and the Automatic Control Laboratory at ETH Zurich. He held a B.Sc. and M.Sc. degree in Electrical Engineering from the University of Tehran and a Ph.D. degree in Management of Technology from EPFL. His primary research interests revolve around low-complexity decision-making, optimization under uncertainty and optimal transport. He is a recipient of the NSF CAREER Award and the SNF Postdoc.Mobility Fellowship.



Gabriel de A. Gleizer is an assistant professor at the Delft Center for Systems and Control (DCSC), Delft University of Technology (The Netherlands). He received his Control and Automation Engineering degree in 2010 and his M.Sc. in Electrical Engineering in 2013 from the Federal University of Rio de Janeiro, Brazil, and his Ph.D. in Systems and Control at DCSC, TU Delft. Before his Ph.D., he was a Research Engineer at GE Global Research in Rio de Janeiro, from 2012 to 2017. His main research interests are hybrid systems, fault diagnosis, networked control systems, and formal methods for quantitative analysis and control.



Peyman Mohajerin Esfahani received the B.Sc. and M.Sc. degrees from Sharif University of Technology, Iran, and the Ph.D. degree from ETH Zurich, Switzerland. He is currently an Associate Professor at the University of Toronto and Delft University of Technology. He joined TU Delft in October 2016, and prior to that, he held several research appointments at EPFL, ETH Zurich, and MIT between 2014 and 2016. His research interests include theoretical and practical aspects of decision-making problems in uncertain and dynamic environments, with applications in engineering systems and management.

He currently serves as an associate editor of *Operations Research*, *Mathematics of Operations Research*, *Mathematical Programming*, *Transactions on Automatic Control*, and *Open Journal of Mathematical Optimization*. He was one of the three finalists for the Young Researcher Prize in Continuous Optimization awarded by the Mathematical Optimization Society in 2016, and a recipient of the 2016 George S. Axelby Outstanding Paper Award from the IEEE Control Systems Society. He received the ERC Starting Grant and the INFORMS Frederick W. Lanchester Prize in 2020. He is the recipient of the 2022 European Control Award.