

A Saddle Point Algorithm for Robust Data-Driven Factor Model Problems

Shabnam Khodakaramzadeh¹, Soroosh Shafiee², Gabriel de Albuquerque Gleizer¹, and Peyman Mohajerin Esfahani^{1,3}

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 general 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.

Keywords: Factor model, covariance matrix estimation, first-order algorithms, dimension reduction, saddle-point problem, robust optimization

1. INTRODUCTION

The correlation structure among observed random variables can be understood using factor analysis in many applications. This correlation is expressed in terms of a smaller number of common *factors* [40], i.e., the information contained in a high-dimensional data vector can be compressed into a small number of common unobserved factors [63]. Mathematically, a high-dimensional data point $\xi \in \mathbb{R}^n$ is represented as the sum of two independent unobserved parts

$$\xi = \Phi\alpha + \omega, \tag{1}$$

where $\Phi \in \mathbb{R}^{n \times r}$ is a tall, full-rank matrix ($n \gg r$), known as *factor loading matrix*, and $\alpha \in \mathbb{R}^r$ is the vector of independent unobserved (latent) factors. Therefore, $\Phi\alpha$, having interrelated components, specifies the low-dimensional representation of ξ . The second part in (1) denoted by ω is known as *idiosyncratic noise*, and has independent components and represents the remaining part of the description [2, 20]. Such low-dimensional decompositions have also been studied in other contexts, such as principal component analysis (PCA). However, PCA is particularly well-suited for scenarios where the data is corrupted by small, unstructured noise [11]. In contrast, the factor model addresses the opposite setting where the noise may be substantial but exhibits a structure, for instance, it has independent components. Another similar problem is compressed sensing, where the noise is still small and unstructured, but r factors are contained in a sparse tall vector α [27].

Date: June 12, 2025.

The authors are with (1) Delft University of Technology, (2) Cornell University, and (3) the University of Toronto. This research is supported by the ERC Starting Grant TRUST-949796.

Originating from the analysis of mental test scores [2, 54], the factor model has found wide-ranging applications across various domains, including control engineering [15], system identification [38, 1, 30], fault detection [60, 62], anomaly detection [61], econometrics [32, 33, 29, 57, 23], and statistics [31, 20]. In the systems and control community, the factor model is particularly valuable for modeling and identifying large-scale or complex networked systems that are subject to uncertainty and disturbances [51]. In the context of machine learning, it serves as an effective tool for dimensionality reduction and feature extraction, leveraging the low-dimensional representation given in (1). We note that factorable structures have also been leveraged in various contexts, such as Markov decision processes [16], robust optimization [22], and matrix factorization problems [42], among many others. This paper aligns most closely with the matrix factorization literature.

1.1. Problem formulation

When the covariance of ξ , denoted by Σ , is available, and assuming that both random variables ξ and ω are zero-mean and independent from each other, we can rewrite (1) as the matrix decomposition

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

where $L = \Phi \Sigma_\alpha \Phi^\top$ is a low rank matrix with $\text{rank}(L) = r$ (i.e., number of factors) in which $\Sigma_\alpha = I$ is the covariance matrix of α , and D is the covariance of noise inheriting the properties of the noise, e.g., it is a non-negative diagonal matrix. In practice, the covariance data Σ is often not available and is only observable through a finite dataset $\{\xi_k\}_{k=1}^N$. Considering this limitation, one may approximate the covariance matrix with its empirical counterpart

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

Under the assumption that the random vector ξ is zero-mean ($\mathbb{E}(\xi) = 0$), one can also consider the data-driven covariance approximation where $\hat{\mu} = 0$ in (3). To robustify to this approximation error, a common practice is to consider a family of covariance matrices in the vicinity of $\hat{\Sigma}$ defined as

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

where d is a generic distance function in the space of matrices, and ε is the radius (size) of the set. Considering the target decomposition (2) and the uncertainty set (4), our *robust data-driven factor model* problem can be formulated as the optimization problem

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

where the cone of PSD matrices \mathbb{S}_+ and the cone of nonnegative diagonal matrices \mathbb{D}_+ capture the structural properties of the decomposition, and the trace operator in the objective function is a standard convexification for the nonconvex rank function [52]. We note that the second line constraint in (5) is indeed equivalent to $d(L + D, \hat{\Sigma}) \leq \varepsilon$, thanks to the conic structural information in the first two constraints. Moreover, the radius ε of the uncertainty ball is a hyperparameter that is determined based on the precision of our data-driven approximation $\hat{\Sigma}$.

1.2. Existing work

Traditionally, the factor model problem was addressed under the assumption that the sample covariance matrix $\widehat{\Sigma}$ is an accurate estimate of the true covariance matrix Σ (i.e., $\Sigma = \widehat{\Sigma}$), which corresponds to $\varepsilon = 0$. Under this assumption, several solution methods based on expectation-maximization algorithm [25, 45], rank-constrained optimization, and asymptotic principal components method [24, 11, 4] have been proposed. However, these approaches may suffer from the inaccuracy of the estimate $\widehat{\Sigma}$, a situation commonly encountered in practice. Consequently, recent research has increasingly focused on versions of the factor model that explicitly account for uncertainty in $\widehat{\Sigma}$, i.e., considering $\varepsilon > 0$. For example, the authors in [19] propose a coordinate descent-type algorithm that minimizes the squared Frobenius norm of residual ($\|\widehat{\Sigma} - L - D\|_F^2$), while [9] develops an algorithm based on conditional gradient method targeting the general Schatten q -norm of this residual ($\|\widehat{\Sigma} - L - D\|_q^q$). The robust factor model problem (5) has also previously been studied in [20] for the special case where the distance function d is the Kullback-Leibler divergence.

While existing approaches are typically designed for a specific choice of distance function d , this study introduces a generic algorithm that only requires access to the linear minimization oracle (LMO)

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

for any symmetric matrix Λ , where the ball $\mathbb{B}_{\varepsilon}^d(\widehat{\Sigma})$ is defined as in (4). An important computational advantage of the proposed algorithm lies in the reduced dimensionality of the linear minimization oracle (LMO) (6). Specifically, the number of decision variables in the LMO (i.e., the matrix Σ) is essentially half that of the original problem (5), which involves both D and L . This reduction becomes particularly beneficial in high-dimensional settings, where solving the LMO is significantly more efficient than addressing the full optimization problem directly. Although the LMO is often tractable and admits near-analytic solutions, the conic constraints associated with the feasible sets \mathbb{S}_+ and \mathbb{D}_+ can make the solutions to (5) computationally demanding.

Finally, we wish to note that this study, along with the literature mentioned above, focuses on a static formulation of the factor model, where the latent factors remain constant over time. However, a dynamic counterpart also exists, as explored in works such as [51, 41, 39], where the temporal evolution of factors is modeled via state-space representations. The study of dynamic factor models, however, lies beyond the scope of this work.

1.3. Main contribution

Given the above literature, we summarize the contributions of this work as follows.

- **Saddle point characterization.** For a general class of distance functions and possible convex conic structural information about the covariance matrix, we reformulate the factor model (5) as a saddle point problem (Proposition 2.1).
- **First-order algorithm & convergence.** Leveraging the saddle point reformulation and given a linear minimization oracle (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, as opposed to the standard sublinear rate, which is enabled by Dykstra’s projection technique (Proposition 2.5).

- **Special LMOs: closed-form description & Lipschitz constant.** The choice of distance function directly influences the corresponding oracle and Lipschitz constant of the dual function, both of which play a central role in the implementation of the proposed algorithm and its convergence performance. We derive the closed-form description for the LMO and its respective Lipschitz constant for the special cases of the Frobenius norm (Proposition 3.1), the Kullback-Leibler divergence (Proposition 3.3), and the Gelbrich (aka Wasserstein) distance (Proposition 3.6). Particular emphasis is devoted to providing a tight, explicit characterization of how key parameters, such as the ambiguity set, affect these quantities.

The theoretical results are validated through extensive numerical experiments, demonstrating the performance of the proposed algorithms. To facilitate reproducibility, we provide an open-source MATLAB library available at https://github.com/skhodakaram/Factor_Model.

Organization. Section 2 presents the saddle point reformulation and the first-order algorithm of the factor model. Section 3 studies the oracle of the three special distance functions in which their closed-form solutions and other regularities are characterized. Section 4 numerically validates the theoretical results, and Section 5 concludes with some final remarks and possible research directions.

Notation. For any symmetric square matrix A , $\lambda_{\max}(A)$ denotes its maximum eigenvalue and its trace is represented by $\text{Tr}(A)$. The nuclear norm of A is denoted as $\|A\|_*$. The matrix denoted as $\text{Diag}(A)$ is a diagonal matrix whose diagonal elements are the diagonal elements of matrix A . For any $A, B \in \mathbb{R}^{n \times m}$, $\langle A, B \rangle = \text{Tr}(A^\top B)$ denotes the inner product of A and B . The respective squared Frobenius norm is denoted by $\|A\|_F^2 := \langle A, A \rangle$. In the special case of $a \in \mathbb{R}^n$ vectors (i.e., $m = 1$), the Frobenius norm coincides with the classical Euclidean norm and is denoted by $\|a\|_2$. The operator $\Pi_{\mathbb{A}}[x]$ denotes the orthogonal projection of the vector x onto the set \mathbb{A} . The element-wise inequality between matrices is denoted by $A \geq B$, and the semidefinite counterpart by $A \succeq B$. The positive semidefinite cone (i.e., all symmetric matrices $A \succeq 0$) is denoted by \mathbb{S}_+ . Given a convex cone $\mathbb{C} \in \mathbb{R}^n$, its dual cone is defined by $\mathbb{C}^* := \{x \in \mathbb{R}^n : \langle x, y \rangle \geq 0, \forall y \in \mathbb{C}\}$. The function $\mathbb{1}_{\mathbb{A}}$ denotes the binary-valued indicator function over the set \mathbb{A} , i.e., $\mathbb{1}_{\mathbb{A}}(x) = 1$ if $x \in \mathbb{A}$; otherwise $= 0$.

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*

$$J^* = \max_{\substack{I - \Lambda \in \mathbb{S}_+ \\ -\Lambda \in \mathbb{D}_+^*}} \min_{\Sigma \in \mathbb{B}_\varepsilon^{\mathbb{d}}(\hat{\Sigma})} \langle \Lambda, \Sigma \rangle, \quad (7)$$

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

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

$$\begin{aligned} \max_{\Lambda} \quad & \min_{\Sigma, L, D} \quad \text{Tr}(L) + \langle \Lambda, \Sigma - L - D \rangle \\ \text{s.t.} \quad & L \in \mathbb{S}_+, \quad D \in \mathbb{D}_+ \\ & \Sigma \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma}), \end{aligned} \quad (8)$$

where the strong duality holds thanks to the usual convex-concave property. Λ is symmetric since Σ , L , and D are symmetric. We note that the usual Slater's condition is not required due to the linearity of the dualized constraint [7]. In the dualized program (8), the decision variables L and D 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 [7, 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., $\mathbb{S}_+^* = \mathbb{S}_+$), and computing the dual cone \mathbb{D}_+^* arrives at the desired program (7). \square

It is worth noting that the LMO (6) is used to find the solution to the inner minimization of the primal-dual reformulation in (7) as a function of the decision variable of the outer maximization problem. The max-min formulation (7) facilitates the development of an optimization algorithm to tackle (5) numerically. The inner minimal value, hereafter referred to as the dual function, is

$$g(\Lambda) := \min_{\Sigma \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda, \Sigma \rangle \quad (9)$$

Note that the dual function $g(\Lambda)$ in (9) is indeed the optimal value corresponding to the optimal point of the LMO in (6). The availability of this oracle motivates us to study the application of first-order algorithms on (9). The Lipschitz continuity of the dual function $g(\Lambda)$ in (9) is a critical regularity condition that ensures the success of the optimization 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 [53]. With this in mind, the next result quantifies the Lipschitz constant of the dual function.

Lemma 2.2 (Dual function Lipschitz constant). *The function $g(\Lambda)$, defined in (9), is Lipschitz continuous with the constant \mathcal{L} , i.e.,*

$$|g(\Lambda_1) - g(\Lambda_2)| \leq \mathcal{L} \|\Lambda_1 - \Lambda_2\|_F, \quad \text{where} \quad \mathcal{L} := \max_{\Sigma \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \|\Sigma\|_F. \quad (10)$$

Proof. Using the definition of the dual function in (9), we have

$$g(\Lambda_1) - g(\Lambda_2) = \min_{\Sigma_1 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda_1, \Sigma_1 \rangle - \min_{\Sigma_2 \in \mathbb{B}_\varepsilon^d(\widehat{\Sigma})} \langle \Lambda_2, \Sigma_2 \rangle$$

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

where the 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

2.2. First-order algorithm and Dykstra projection oracle

We propose a first-order algorithm using the LMO (6) to solve problem (7), which is the primal-dual reformulation of the factor model problem.

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

$$\begin{cases} \Sigma_t &= O(\Lambda_t) \\ \Lambda_{t+1} &= \Pi_{\mathbb{S}_1 \cap \mathbb{S}_2} [\Lambda_t + \delta \Sigma_t] \\ \bar{\Lambda}_t &= \frac{t-1}{t} \bar{\Lambda}_{t-1} + \frac{1}{t} \Lambda_t, \quad \bar{\Sigma}_t = O(\bar{\Lambda}_t) \end{cases} \quad (11)$$

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 the decision variable Λ in (7), O is the linear minimization oracle defined in (6), and δ is a constant stepsize. Then, after T iterations of algorithm (11) we have

$$0 \leq \langle \Lambda^*, \Sigma^* \rangle - \langle \bar{\Lambda}_T, \bar{\Sigma}_T \rangle \leq \frac{\|\Lambda_1 - \Lambda^*\|^2}{2\delta T} + \frac{\delta}{2} \mathcal{L}^2 \quad (12)$$

where (Λ^*, Σ^*) is a saddle point solution of (7), Λ_1 is the initial condition of (11), and the constant \mathcal{L} is the Lipschitz constant defined in (10).

Proof. The proof follows from the standard projected subgradient method. To see this, note that our LMO (6) is indeed the subgradient of our dual function (9), i.e., $O(\Lambda) = \frac{\partial g(\Lambda)}{\partial \Lambda}$. This observation is a consequence of Danskin's theorem [6, (A.22), p 154]. Therefore, the first two steps in (11) effectively implement the projected subgradient ascent applied to the dual function d defined in (9) while the dynamics of $\bar{\Lambda}_t$ is the averaging method 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., [8, Proposition 3.2.4]), we have the error bound (12). \square

Remark 2.4 (Diminishing stepsize & averaging). *Minimizing the error bound in (12) with respect to the constant stepsize δ reveals that the optimal choice is $\delta = \mathcal{O}(1/\sqrt{T})$, which yields a suboptimality gap of $\mathcal{O}(1/\sqrt{T})$; see for example [3, Corollary 2]. However, such constant stepsizes often perform poorly in practice. Alternatively, it is shown in [53] that a diminishing anytime stepsize $\delta_t = \mathcal{O}(1/\sqrt{t})$ achieves similar convergence, up to a logarithmic factor. This justifies the widespread use of $\mathcal{O}(1/\sqrt{t})$ stepsizes, which we adopt in our experiments. We further note that other stepsize management approaches, such as averaging techniques, have been employed in the context of the Lagrangian dual for constrained optimization [47] and general saddle-point problems [48].*

It should be noted that the possible difficulty in the proposed algorithm (11) is the projection on the intersection of \mathbb{S}_1 and \mathbb{S}_2 . Depending on the cones of \mathbb{S}_1 and \mathbb{S}_2 , this projection operation may be

difficult. To address this issue, we propose Dykstra's algorithm as an effective approach to implement this operation, which is described in Algorithm 1 [55, 28].

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

- 1: **Input:** $U_2^0 = \Lambda$, $Z_1^0 = 0$, $Z_2^0 = 0$, $k = 1$
 - 2: **while** $\frac{\|U_2^k - U_1^k\|_F}{\|U_2^k\|_F} \geq 10^{-6}$ **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
-

2.3. Analysis of Dykstra's projection algorithm

This section addresses the projection step in algorithm (11). Formally, for a given symmetric matrix $\bar{\Lambda}$, we aim to solve the optimization problem

$$\Lambda^* = \arg \min \left\{ \frac{1}{2} \|\Lambda - \bar{\Lambda}\|_F^2 : \Lambda \in \mathbb{S}_1 \cap \mathbb{S}_2 \right\}. \quad (13)$$

Projecting onto the intersection of convex sets is generally viewed as a convex optimization problem, often called the best approximation problem. Dykstra's projection algorithm, introduced in [28, 17, 37], efficiently solves this problem. Algorithm 1 outlines Dykstra's projection algorithm for problem (13). Unlike the alternating projection algorithm proposed in [58], Dykstra's projection algorithm is guaranteed to converge to the unique solution of the best approximation problem. Specifically for problem (13), given that $\mathbb{S}_1 \cap \mathbb{S}_2 \neq \emptyset$, the sequence generated by this algorithm converges *asymptotically* to the unique solution of (13) due to [17, Theorem 2]

$$\lim_{k \rightarrow \infty} \|\Lambda^k - \Lambda^*\|_F = 0,$$

where Λ is the output of Dykstra's projection algorithm.

Despite its simplicity, Dykstra's projection algorithm requires no additional assumptions for convergence, unlike splitting methods such as Douglas-Rachford splitting, which need further regularity conditions; see [5, Corollary 28.3]. Convergence rate analysis of Dykstra's projection algorithm and its variants has been studied extensively in recent years. In particular, when all sets are polyhedral, [44, 26, 56] establish the linear convergence of the algorithm. Additionally, [34, 37] prove that Dykstra's projection algorithm can be viewed as a dual coordinate gradient descent method. Using this connection and showing that the dual objective function admits a so-called Kurdyka-Łojasiewicz property [12, 13], [59, Theorem 5.3] prove a linear convergence rate of the algorithm for a certain class of conic problems and under some regularity conditions. The following theorem establishes that these conditions are satisfied for certain points $\bar{\Lambda}$ when we solve (13). For a set \mathbb{A} , we denote by $\mathcal{N}_{\mathbb{A}}(x) := \{v : \langle v, y - x \rangle \leq 0, \forall y \in \mathbb{A}\}$ the normal cone of the set at x .

Proposition 2.5 (Linear convergence for projection). *Suppose that the optimizer of the program (13) satisfies $\Lambda^* - \bar{\Lambda} \in \text{rint}(\mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\bar{\Lambda}))$. Then, Algorithm 1 converges linearly to Λ^* .*

Proof. The proof concludes if we show all requirements of [59, Theorem 5.3] specified in [59, Assumption 4.1] are satisfied for problem (13). Observe that \mathbb{S}_1 is a polyhedron, and it is therefore C^2 -cone reducible in the sense of [14, Definition 3.135]. The same also holds for \mathbb{S}_2 as it can be reduced to a (shifted) semidefinite cone. Thus, the first requirement in [59, Assumption 4.1] is satisfied. The second requirement also holds as $\text{rint}(\mathbb{S}_1 \cap \mathbb{S}_2)$ is nonempty. Finally, the third requirement is satisfied due to the assumption we made that $\Lambda^* - \bar{\Lambda} \in \text{rint}(\mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\bar{\Lambda}))$. This concludes the proof. \square

We note that the assumption $\Lambda^* - \bar{\Lambda} \in \text{rint}(\mathcal{N}_{\mathbb{S}_1 \cap \mathbb{S}_2}(\bar{\Lambda}))$ cannot be verified a priori. However, this assumption is both standard in the literature and crucial for establishing linear convergence. For further discussion on its necessity for linear convergence, we refer the reader to [59, Example 5.1].

3. SPECIAL CASES OF LINEAR MINIMIZATION ORACLE

In this section, we study three special cases of the LMO (6) in more detail, in particular in view of their computational complexity and the respective Lipschitz continuity of the dual function (9). This includes (1) the Frobenius distance, (2) the Kullback-Leibler divergence, and (3) the Gelbrich distance.

3.1. Frobenius norm

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

Proposition 3.1 (Frobenius oracle & Lipschitz constant). *Consider the 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*

$$\Sigma^*(\Lambda, \gamma) := \Pi_{\succeq 0} \left[\hat{\Sigma} - \frac{1}{2\gamma} \Lambda \right], \quad (14)$$

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

$$\max_{0 < \gamma \leq \|\Lambda\|_{\text{F}}} \langle \Lambda, \Sigma^*(\Lambda, \gamma) \rangle + \gamma (\|\Sigma^*(\Lambda, \gamma) - \hat{\Sigma}\|_{\text{F}}^2 - \varepsilon^2). \quad (15)$$

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

$$\mathcal{L} \leq \varepsilon + \|\hat{\Sigma}\|_{\text{F}} \quad (16)$$

Proof. Let us start with part (i). Introducing the Lagrange multiplier γ corresponding to the constraint $\|\Sigma - \hat{\Sigma}\|_{\text{F}}^2 \leq \varepsilon^2$, the dual program of the oracle optimization (6) is

$$\max_{\gamma \geq 0} \min_{\Sigma \succeq 0} \langle \Lambda, \Sigma \rangle + \gamma (\|\Sigma - \hat{\Sigma}\|_{\text{F}}^2 - \varepsilon^2). \quad (17)$$

To compute the inner minimizer explicitly, we note that the Hessian of the Lagrangian function in (17) with respect to Σ is a scaled identity. This allows us to first compute the optimal solution of the unconstrained problem (ignoring the positivity constraint $\Sigma \succeq 0$), and then project that onto the PSD

cone. With this in mind, the first-order optimality condition for the unconstrained program of the inner minimization yields $\Lambda + 2\gamma(\Sigma^* - \widehat{\Sigma}) = 0$, for every (Λ, γ) . Notice that the inner optimization in (17) has a unique solution described by this linear algebraic equation, and as such, it explains the closed-form projected solution $\Sigma^*(\Lambda, \gamma)$ in (14). Substituting $\Sigma^*(\Lambda, \gamma)$ in the inner minimization of (17) arrives at the program (15) whose solution γ^* together with the closed-form solution $\Sigma^*(\Lambda, \gamma^*)$ determines the saddle point of the max-min problem (17). It is important to note that excluding $\gamma = 0$ is without loss of generality since the objective function of the inner minimization in (17) is lower-semicontinuous in γ (pointwise minimum of a continuous function in both variables (γ, Λ)). The final step of the proof in the first part is to show that the optimal γ^* is bounded by $\|\Lambda\|_F$. To this end, we use [47, Lemma 1], which relies on the existence of a Slater point. Considering $\widehat{\Sigma}$ as the Slater point, [47, Lemma 1] offers the upper bound

$$\gamma^* \leq \frac{1}{\varepsilon} \left(\langle \widehat{\Sigma}, \Lambda \rangle - \min_{\Sigma \in \mathbb{B}_\varepsilon^F(\widehat{\Sigma})} \langle \Sigma, \Lambda \rangle \right) = \max_{\Sigma \in \mathbb{B}_\varepsilon^F(\widehat{\Sigma})} \frac{1}{\varepsilon} \langle \widehat{\Sigma} - \Sigma, \Lambda \rangle \leq \|\Lambda\|_F \quad (18)$$

where the last inequality is a direct consequence of lifting the positivity constraint $\Sigma \succeq 0$ and the fact that the Frobenius norm is self-dual. This concludes the proof of part (i).

Next, we continue with the proof of part (ii). Thanks to Lemma 2.2, in particular (10), the Lipschitz constant \mathcal{L} satisfies

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

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

Proposition 3.1(i) provides an efficient computational way to implement the LMO (6) since the scalar concave maximization problem (15) can be efficiently solved through bisection within the feasible region $\gamma \in (0, \|\Lambda\|_F]$.

3.2. Kullback-Leibler divergence

Our second case study is the Kullback-Leibler (KL) divergence, also known as *relative entropy*, between two normal distributions with the same means and the covariance matrices Σ and $\widehat{\Sigma}$ [21, Chap. 2]. To make this definition well defined, we assume throughout this section that $\widehat{\Sigma} \succ 0$, and as such invertible. Moreover, we assume that the vectors ξ , α , and ω are Gaussian random vectors. The result can be extended to the case of rank deficient $\widehat{\Sigma}$ in which the KL-ball with finite radius is also required to carry the same null space. Nonetheless, for the sake of clarity, we do not pursue this level of generality here.

Definition 3.2 (Kullback-Leibler divergence). *The Kullback-Leibler divergence between two zero-mean normal distributions with covariance matrices Σ and $\widehat{\Sigma}$ is*

$$\text{KL}(\Sigma || \widehat{\Sigma}) = \frac{1}{2} \left(-\log \det \Sigma + \log \det \widehat{\Sigma} + \text{Tr} \left(\Sigma \widehat{\Sigma}^{-1} \right) - n \right).$$

The KL divergence is a popular statistical distance with a rich geometric interpretation that finds various applications in control, information theory, economics, and finance, to name a few. While the KL divergence is a similarity measure, unlike the Frobenius norm, it is not symmetric in its arguments

and does not satisfy the triangle inequality. Next, we provide an explicit description of its respective oracle and dual function Lipschitz constant.

Proposition 3.3 (KL oracle & Lipschitz constant). *Consider the LMO (6) for a given $\widehat{\Sigma}$ where the distance function is the KL divergence $d(\Sigma, \widehat{\Sigma}) = \text{KL}(\Sigma || \widehat{\Sigma})$*

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

$$\Sigma^*(\Lambda, \gamma) := \left(\widehat{\Sigma}^{-1} + \frac{2}{\gamma} \Lambda \right)^{-1}. \quad (19a)$$

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

$$\begin{cases} \text{KL}(\Sigma^*(\Lambda, \gamma^*) || \widehat{\Sigma}) - \varepsilon = 0 \\ \max \left\{ 0, 2\lambda_{\max}(-\widehat{\Sigma}^{\frac{1}{2}} \Lambda \widehat{\Sigma}^{\frac{1}{2}}) \right\} < \gamma^* \leq \|\widehat{\Sigma}^{\frac{1}{2}} \Lambda \widehat{\Sigma}^{\frac{1}{2}}\|_* \left(\sqrt{\frac{6}{\varepsilon}} \mathbb{1}_{[0, \frac{1}{24}]} + \left(6 + \frac{1}{4\varepsilon}\right) \mathbb{1}_{(\frac{1}{24}, \infty)} \right) \end{cases} \quad (19b)$$

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

$$\mathcal{L} \leq \left(n\sqrt{6\varepsilon} \mathbb{1}_{[0, \frac{1}{24}]} + n\left(6\varepsilon + \frac{1}{4}\right) \mathbb{1}_{(\frac{1}{24}, \infty)} + 1 \right) \|\widehat{\Sigma}\|_{\text{F}}. \quad (20)$$

We note that the authors of [20] also analyze a related closed-form solution to the dual of the KL oracle (19a) (cf. [20, Proposition 4.1]). However, our formulation refines both the closed-form solution (19a) and, more specifically, the characterization of the feasible region for the optimal dual multiplier (19b). The proof of Proposition 3.3 builds on the following technical lemma, which characterizes a lower bound for the KL divergence and, as such, a subset of the KL-ball in terms of the eigenvalues of the geometric average of the matrices involved.

Lemma 3.4 (KL-lower bound). *The KL divergence $\text{KL}(\Sigma || \widehat{\Sigma})$ in Definition 3.2 satisfies*

$$2 \text{KL}(\Sigma || \widehat{\Sigma}) \geq \sum_{i=1}^n f(\lambda_i(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}})) \quad \text{where} \quad f(\lambda) := \frac{1}{3}(\lambda - 1)^2 \mathbb{1}_{[0, \frac{3}{2}]} + \left(\frac{1}{3}\lambda - \frac{5}{12}\right) \mathbb{1}_{(\frac{3}{2}, \infty)}. \quad (21a)$$

In particular, for any radius $\varepsilon \geq 0$, if $\text{KL}(\Sigma || \widehat{\Sigma}) \leq \varepsilon$ we then have

$$|\lambda_{\max}(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) - 1| \leq \sqrt{6\varepsilon} \mathbb{1}_{[0, \frac{1}{24}]} + \left(6\varepsilon + \frac{1}{4}\right) \mathbb{1}_{(\frac{1}{24}, \infty)}. \quad (21b)$$

Proof. Following Definition 3.2, we have

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

where the equalities above hold due to the symmetric property of the trace and determinant operators (i.e., $\text{Tr}(AB) = \text{Tr}(BA)$ and $\det(AB) = \det(BA)$), and the inequality follows from the Taylor series expansion with degree 2 of the convex function $\lambda - 1 - \log(\lambda)$ at point $\lambda = 1$ within the interval $[0, \frac{3}{2}]$, followed by a linear extension of this lower bound for $\lambda > \frac{3}{2}$. This concludes the assertion (21a). Concerning the second part, it suffices to use an inverse function argument for the lower bound function f in (21a) and apply the simple inequality $\max_{i \leq n} f(\lambda_i) \leq \sum_{i \leq n} f(\lambda_i)$. This yields the desired assertion in (21b). \square

The lower bound in Lemma 3.4 provides a means to introduce a superset for the KL-ball, which is useful for bounding the closed-form description of the respective oracle and its Lipschitz constant of the optimal dual function with respect to the ball radius.

Proof of Proposition 3.3. The proof of both parts follows similar lines as in Proposition 3.1. Concerning (i), we note that dualizing the distance function constraint corresponding to $\text{KL}(\Sigma||\widehat{\Sigma}) \leq \varepsilon$ in the LMO (6) yields

$$\max_{\gamma \geq 0} \min_{\Sigma \succeq 0} \langle \Lambda, \Sigma \rangle + \frac{\gamma}{2} \left(-\log \det \Sigma + \log \det \widehat{\Sigma} + \text{Tr} \left(\Sigma \widehat{\Sigma}^{-1} \right) - n \right) - \gamma \varepsilon. \quad (22)$$

When $\gamma > 0$, we can exclude the boundary of the PSD cone for the inner optimizer Σ (i.e., $\Sigma \succ 0$) due to the term $-\log \det \Sigma$ in the objective (22). Similar to the proof of Proposition 3.1, we can also exclude the case of $\gamma = 0$ since the optimal value of the inner minimization is lower-semicontinuous in γ . Therefore, the first-order optimality condition for the inner minimization problem yields the unique inner minimizer $\Sigma^*(\Lambda, \gamma)$ in (19a). Substituting $\Sigma^*(\Lambda, \gamma)$ in (22), we arrive at

$$\max_{\gamma > 0} \langle \Lambda, \Sigma^*(\Lambda, \gamma) \rangle + \frac{\gamma}{2} \left(-\log \det \Sigma^*(\Lambda, \gamma) + \log \det \widehat{\Sigma} + \text{Tr} \left(\Sigma^*(\Lambda, \gamma) \widehat{\Sigma}^{-1} \right) - n \right) - \gamma \varepsilon. \quad (23)$$

Applying the first-order optimality condition to the optimization (23) yields the algebraic equation (19b). Next, we derive the lower and upper bounds for the optimal solution γ^* solving (23) (or equivalently the algebraic equation (19b)). For the lower bound of the optimizer γ^* in (19b), note that the inner optimizer $\Sigma^*(\Lambda, \gamma^*) \succ 0$, and as such, we have the lower bound $\widehat{\Sigma}^{-1} + \frac{2}{\gamma^*} \Lambda \succ 0$, which implies $\gamma^* > 2\lambda_{\max}(-\widehat{\Sigma}^{\frac{1}{2}} \Lambda \widehat{\Sigma}^{\frac{1}{2}})$. The above bound, together with the original non-negativity constraint, concludes the lower bound.

The last part of the proof is to show the upper bound of the optimizer γ^* in (19b). The analysis follows similar lines as in Proposition 3.1, starting from the corresponding bound (18), that is,

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

where the last inequality follows from (21b) in Lemma 3.4 in which the radius is 2ε . Note further that the objective function of the above program can be upper bounded by

$$\langle \widehat{\Sigma} - \Sigma, \Lambda \rangle = \langle I - \widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}, \widehat{\Sigma}^{\frac{1}{2}} \Lambda \widehat{\Sigma}^{\frac{1}{2}} \rangle \leq \max_{i \leq n} |\lambda_i(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}) - 1| \|\widehat{\Sigma}^{\frac{1}{2}} \Lambda \widehat{\Sigma}^{\frac{1}{2}}\|_*, \quad (25)$$

where $\|\cdot\|_*$ is the nuclear matrix norm. Considering the above upper bound in the program (24) and replacing λ_{\max} yields the desired upper bound for γ^* in (19b).

Regarding the Lipschitz constant in part (ii), following the bound (10), we have

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

where the last inequality above follows from (21b) in Lemma 3.4. \square

3.3. Gelbrich distance

Our final case study examines the Gelbrich distance, denoted by $G(\Sigma, \widehat{\Sigma})$, which measures the distance between two distributions with identical means and covariance matrices Σ and $\widehat{\Sigma}$. While a more general form of this distance exists for distributions with different means, we omit it here as it is not necessary for our study [36].

Definition 3.5 (Gelbrich distance). *The Gelbrich distance between two zero-mean distributions with the positive semidefinite covariance matrices Σ and $\widehat{\Sigma}$ is*

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

It is well known that the Gelbrich distance is a lower bound for the Wasserstein distance between two distributions, and its maximal coincides with the Wasserstein distance for the subclass of elliptical distributions (e.g., normal distributions [35, Theorem 2.1]). The LMO (6) under the Gelbrich distance in Definition 3.5 has been studied before [49, 50]. In parallel with the other two oracles discussed in the preceding sections, we adapt the Gelbrich LMO from these studies, incorporating a slight generalization that extends to arbitrary matrices (rather than only positive semidefinite matrices as in [50]). Additionally, we provide a tight upper bound for the univariate optimization problem and establish the Lipschitz constant of the corresponding dual function, the two parameters contributing to the convergence of the proposed algorithm.

Proposition 3.6 (Gelbrich oracle & Lipschitz constant). *Consider the LMO (6) for a given $\widehat{\Sigma} \succeq 0$ 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*

$$\Sigma^*(\Lambda, \gamma) := \gamma^2(\gamma I + \Lambda)^{-1} \widehat{\Sigma} (\gamma I + \Lambda)^{-1}. \quad (26a)$$

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

$$\begin{cases} \max_{\gamma \geq 0} & \gamma(\varepsilon^2 + \langle \gamma(\gamma I + \Lambda)^{-1} - I, \widehat{\Sigma} \rangle) \\ \text{s.t.} & \lambda_{\max}(-\Lambda) < \gamma \leq \|\Lambda\|_{\text{F}} (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon) \end{cases} \quad (26b)$$

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

$$\mathcal{L} \leq (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon)\varepsilon + \|\widehat{\Sigma}\|_{\text{F}} \quad (27)$$

We note that the quasi-closed form description (26a) and the lower bound of the dual multiplier in (26b) were previously proposed in [50, Proposition A.2]. However, when the matrix Λ is indefinite, which is an important case in the factor model in this study, the upper bound of the dual multiplier in (26b) and the Lipschitz constant of the dual function in (27) are first introduced here. The proof of Proposition 3.6 builds on the following technical lemma that characterizes the regularity properties of the Gelbrich distance and the respective ball.

Lemma 3.7 (Gelbrich lower bound). *The Gelbrich distance in Definition 3.5 satisfies*

$$G(\Sigma, \widehat{\Sigma}) \geq \max \left\{ \left\| \lambda^{\frac{1}{2}}(\Sigma) - \lambda^{\frac{1}{2}}(\widehat{\Sigma}) \right\|_2, \lambda_{\max}^{-\frac{1}{2}}(\Sigma + \widehat{\Sigma} + 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \|\Sigma - \widehat{\Sigma}\|_{\text{F}} \right\}. \quad (28a)$$

In particular, for any radius $\varepsilon \geq 0$, if $G(\Sigma, \widehat{\Sigma}) \leq \varepsilon$ we then have

$$G(\Sigma, \widehat{\Sigma}) \geq \frac{\|\Sigma - \widehat{\Sigma}\|_F}{2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon}, \quad (28b)$$

which essentially implies that the squared Gelbrich distance is strongly convex with respect to the Frobenius norm, uniformly over any compact set of the positive semidefinite cone.

Proof. We first start with the first term on the right-hand side of the inequality (28a). Note that

$$\begin{aligned} G^2(\Sigma, \widehat{\Sigma}) &= \text{Tr} \left(\Sigma + \widehat{\Sigma} - 2 \left(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) = \sum_{i \leq n} \lambda_i(\Sigma) + \lambda_i(\widehat{\Sigma}) - 2\lambda_i^{\frac{1}{2}}(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}) \\ &\geq \sum_{i \leq n} \lambda_i(\Sigma) + \lambda_i(\widehat{\Sigma}) - 2\lambda_i^{\frac{1}{2}}(\widehat{\Sigma})\lambda_i^{\frac{1}{2}}(\Sigma) = \sum_{i \leq n} (\lambda_i^{\frac{1}{2}}(\Sigma) - \lambda_i^{\frac{1}{2}}(\widehat{\Sigma}))^2 = \|\lambda^{\frac{1}{2}}(\Sigma) - \lambda^{\frac{1}{2}}(\widehat{\Sigma})\|_2^2, \end{aligned} \quad (29)$$

where the above inequality follows from the matrix version of the Hardy-Littlewood-Polya inequality [18, Theorem 3.2]. With regard to the second term on the right-hand side of (28a), note that

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

where we use the symmetric property $\text{Tr}(AB) = \text{Tr}(BA)$ to derive the first and second equalities. Since $\text{Tr}(AB) \geq 0$ for any $A, B \succeq 0$, the above equality leads to

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

which, together with (29), concludes the inequality (28a). To derive the inequality (28b), we use the second part on the right-hand side of (28a) when the term $\lambda_{\max}(\Sigma + \widehat{\Sigma} + 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}})$ is upper bounded over the Gelbrich ball with radius ε . To this end, note that

$$\varepsilon \geq G(\Sigma, \widehat{\Sigma}) \geq \|\lambda^{\frac{1}{2}}(\Sigma) - \lambda^{\frac{1}{2}}(\widehat{\Sigma})\|_2 \geq |\lambda_{\max}^{\frac{1}{2}}(\Sigma) - \lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma})|.$$

The above inequality implies that

$$G(\Sigma, \widehat{\Sigma}) \leq \varepsilon \quad \Rightarrow \quad \lambda_{\max}^{\frac{1}{2}}(\Sigma) \leq \lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon. \quad (30)$$

With the above observation, we can deduce that

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

where, the first and second inequalities are the result of sub-additivity and sub-multiplicativity of the maximum eigenvalue of the positive semidefinite cone (i.e., $\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$, and $\lambda_{\max}(AB) \leq \lambda_{\max}(A)\lambda_{\max}(B)$, for all $A, B \succeq 0$). It now suffices to apply the above bound in (28a) to arrive at the bound (28b). \square

Remark 3.8 (Gelbrich strong convexity). *The inequality (28b) in Lemma 3.7 suggests that the squared Gelbrich distance is strongly convex with respect to the Frobenius norm, a property which is of interest in view of optimization algorithms. In comparison with the earlier results in the literature [10, Theorem 1], the strong convexity relation (28b) does not depend on the minimum eigenvalue of the reference matrix $\widehat{\Sigma}$, making it particularly useful when the ball contains low-rank matrices.*

Proof of Proposition 3.6. With regards to the quasi-closed form description (26a) and the lower bound of the dual multiplier in (26b), we refer to parts (i) and (ii) in [50, Proposition A.2]. Concerning the upper bound (26b), we follow the approach used for the Frobenius distance and Kullback-Leibler divergence and derive a similar upper bound to (18) as

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

where the first inequality in the last line is the application of the Cauchy-Schwarz inequality to the objective function and the bound (28b) in Lemma 3.7 for the Gelbrich distance in the constraint. This yields the desired upper bound (26b). Regarding the Lipschitz constant (27), we again apply the upper bound (28b) from Lemma 3.7 to the Lipschitz constant (10) from Lemma 2.2 and arrive at

$$\mathcal{L} = \max_{\Sigma \in \mathbb{B}_\varepsilon^G(\widehat{\Sigma})} \|\Sigma\|_F \leq \max_{\Sigma \in \mathbb{B}_\varepsilon^G(\widehat{\Sigma})} \|\Sigma - \widehat{\Sigma}\|_F + \|\widehat{\Sigma}\|_F \leq (2\lambda_{\max}^{\frac{1}{2}}(\widehat{\Sigma}) + \varepsilon)\varepsilon + \|\widehat{\Sigma}\|_F,$$

which concludes (27). \square

4. NUMERICAL EXAMPLE

We pursue two main objectives through our numerical investigations here: (i) assessment of the performance of our proposed algorithm to solve the factor model (5), particularly in comparison with the existing off-the-shelf solver MOSEK, and (ii) observing the effect of the hyperparameter ε on the accuracy of the covariance matrix estimation. For this purpose, we present the results of our numerical investigation on the convergence of the proposed algorithm, observation of the estimation error of Σ_{True} , and comparison between the computation time of our proposed algorithm and MOSEK. The stopping condition used for the algorithm (11) is the normalized *relative* change defined as

$$\frac{|\langle \Lambda_t, \Sigma_t \rangle - \langle \Lambda_{t-1}, \Sigma_{t-1} \rangle|}{|\langle \Lambda_t, \Sigma_t \rangle|} \leq 10^{-6}. \quad (31)$$

4.1. Synthetic data generation

Every implementation consists of two steps: First, selecting the ground-truth covariance matrix with the decomposition $\Sigma_{\text{True}} = D_{\text{True}} + L_{\text{True}}$, and then producing an approximation of Σ_{True} , we use noisy independently identically distributed (iid) samples of $\{\xi_k\}_{k=1}^N$ based on relation (1) and compute their empirical covariance matrix $\widehat{\Sigma}$ in (3). These two steps are described in more detail as follows.

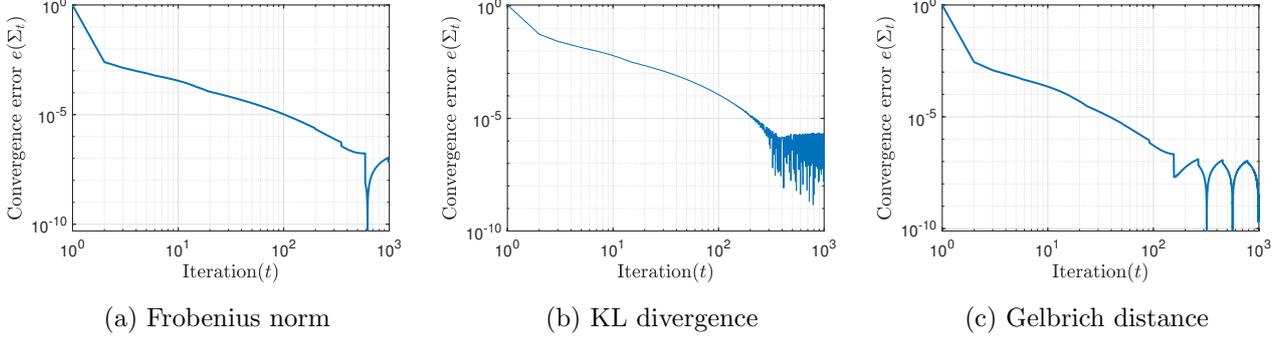


FIGURE 1. Convergence error (32) of algorithm (11) with the projection Algorithm 1

Ground-truth covariance matrix generation: A pseudorandom tall matrix $\Phi_{\text{True}} \in \mathbb{R}^{n \times r}$, and a pseudorandom diagonal PSD matrix $D_{\text{True}} \in \mathbb{R}^{n \times n}$ which are the true factor loading matrix and the true covariance matrix of idiosyncratic noise are created using *rand* function in MATLAB with *rng(1, 'twister')* and *rng(0)*, respectively. The r -dimensional covariance matrix Σ_α is equal to the identity matrix. The matrix Σ_{True} is calculated based on (2), using $L_{\text{True}} = \Phi_{\text{True}} \Phi_{\text{True}}^\top$ and D_{True} matrices. It is worth noting that the matrices Φ_{True} and D_{True} are adjusted to be far away from the origin. To be more precise, the minimum element of these matrices is lower bounded by a certain amount (in our implementations, this is 5, by adding 5 to the *rand* function) so that the resulting Σ_{True} matrix is far enough from the origin, to exclude the trivial solution $\Sigma = 0$, when $\Lambda \succeq 0$ and the ball includes the origin.

Sample generation: The number of samples is selected as $N = 15n$ and the factors samples $\{\alpha_k\}_{k=1}^N$ are pseudorandomly generated from a normal distribution with mean zero and the covariance matrix $\Sigma_\alpha = I$. Similarly, using a normal distribution with mean zero and the covariance matrix equal to D_{True} , the noise samples $\{\omega_k\}_{k=1}^N$ are produced. Then, the samples of observed vectors are $\xi_k = \Phi_{\text{True}} \alpha_k + \omega_k$ for all $k \in 1, \dots, N$, based on relation (1).

4.2. Convergence

The primal-dual reformulation of the factor model problem (7) is solved using Algorithm (11) with diminishing stepsize ($\delta_t = \frac{1}{\sqrt{t}}$) for three cases in which $d(\Sigma, \hat{\Sigma})$ is defined based on $F(\Sigma, \hat{\Sigma})$, $\text{KL}(\Sigma || \hat{\Sigma})$, and $G(\Sigma, \hat{\Sigma})$, respectively. These algorithms are executed with $\varepsilon = 1$ and 10^4 iterations. We note that 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 convergence of the algorithm with respect to this value by computing the normalized convergence error

$$e(\Sigma_t) := \frac{|\langle \Lambda_t, \Sigma_t \rangle - \langle \Lambda^*, \Sigma^* \rangle|}{|\langle \Lambda^*, \Sigma^* \rangle|}, \quad (32)$$

which reflects each iteration's objective value's normalized error with respect to the optimal objective value. The values of the parameters are chosen as $n = 20$ and $r = 4$. The results of this experiment are reported in Figure 1. We emphasize that these simulations are executed without considering the stopping condition (31), to prevent them from stopping earlier than 10^4 iterations. As depicted in Figure 1, after 100 iterations, the normalized error (32) decreases to around 0.001%, 0.01%, and $5 \times 10^{-5}\%$ for

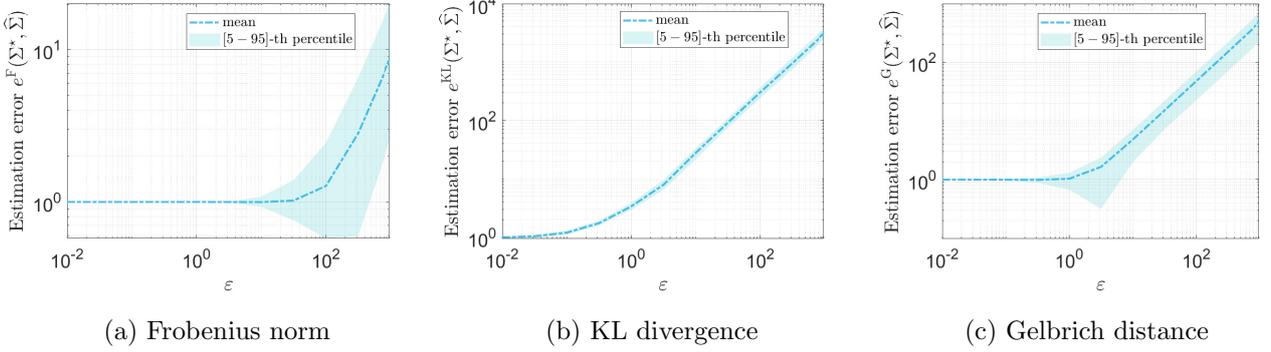


FIGURE 2. Estimation error (34) of the ground-truth Σ_{True}

the cases where $d(\Sigma, \hat{\Sigma})$ is $F(\Sigma, \hat{\Sigma})$, $\text{KL}(\Sigma || \hat{\Sigma})$, and $G(\Sigma, \hat{\Sigma})$, respectively. The performance observed here validates the theoretical convergence result of Proposition 2.3 for all the distance functions.

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

In this numerical study, the effect of the hyperparameter ε on the estimation error of the true covariance matrix Σ_{True} is investigated for N_{exp} experiments. In each experiment, the factor model problem is solved using Algorithm (11) with diminishing stepsize ($\delta_t = \frac{1}{\sqrt{t}}$) for all of the values of ε in the set Ω as defined in

$$\Omega = \{0.01(\sqrt{10})^i \mid i = 0, 1, \dots, 10\} \quad (33)$$

In all experiments, for these values of ε , the algorithm (11) is executed for maximum $t_{\text{end}} = 10^4$ iterations, while considering the stopping criterion (31). The data is generated through the synthetic data generation procedure in Section 4.1 with fixed true covariance matrices D_{True} and Σ_{True} for all of the experiments. We note that in each experiment, the same samples of observed vectors (ξ_k , $k = 1, \dots, N$) are used to solve the problem for different values of ε , while these samples vary among different experiments. The values of the parameters that are used in this simulation are chosen as $n = 20$, $r = 4$, and $N_{\text{exp}} = 100$. For numerical illustration, we consider the normalized estimation error of Σ_{True} defined as

$$e^d(\Sigma^*, \hat{\Sigma}) := \frac{d(\Sigma^*, \Sigma_{\text{True}})}{d(\hat{\Sigma}, \Sigma_{\text{True}})}. \quad (34)$$

The simulation results consist of the mean and [5 – 95]-th percentile of the estimation error of Σ_{True} corresponding to the various values of ε for all of the N_{exp} experiments, which are executed for each of the three special cases discussed in Section 3.

The numerical results of this experiment are reported in Figure 2. A sweet spot is observed for $\varepsilon = 100$ and $\varepsilon = \sqrt{10}$ corresponding to $F(\Sigma, \hat{\Sigma})$ and $G(\Sigma, \hat{\Sigma})$, respectively, indicating that the factor model improves the estimation of Σ_{True} in comparison to $\hat{\Sigma}$ defined in (3), for different cases of $d(\Sigma, \hat{\Sigma})$ corresponding to $F(\Sigma, \hat{\Sigma})$ and $G(\Sigma, \hat{\Sigma})$. To be more precise, based on the simulation results, in 59% of the experiments corresponding to $F(\Sigma, \hat{\Sigma})$, and in 56% of the experiments corresponding to $G(\Sigma, \hat{\Sigma})$, an improvement in the estimation of Σ_{True} , compared to $\hat{\Sigma}$, is observed. We note that for the case of

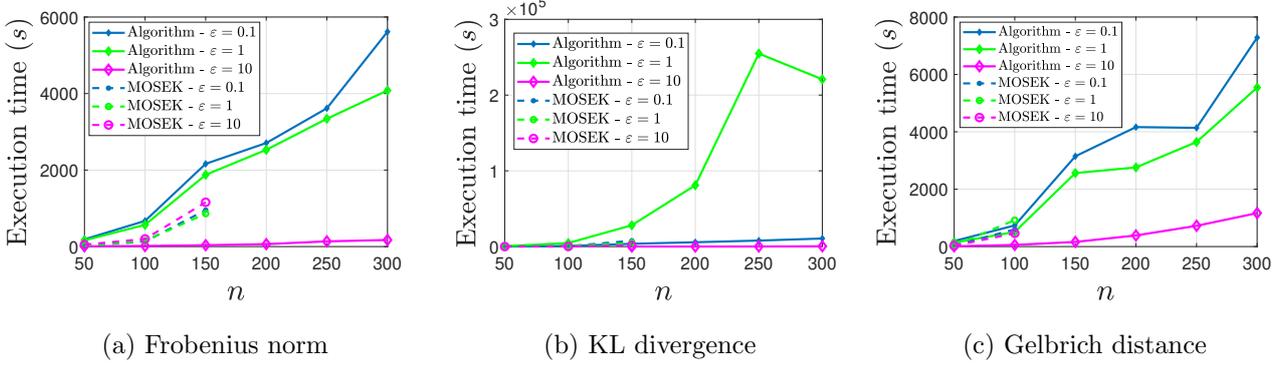


FIGURE 3. Computational time comparison of algorithm (11) and MOSEK

$d(\Sigma, \widehat{\Sigma})$ corresponding to $\text{KL}(\Sigma || \widehat{\Sigma})$, although a distinguishable sweet spot has not been observed, in 43% of the experiments, a slight improvement in the estimation of Σ_{True} , compared to $\widehat{\Sigma}$, is observed.

4.4. Execution time

In this subsection, the execution time of the proposed first-order algorithm is investigated and compared with the off-the-shelf solver MOSEK. Algorithm (11) with diminishing stepsize ($\frac{1}{\sqrt{t}}$) is used for these simulations with $\varepsilon \in \{0.1, 1, 10\}$, considering the stopping condition (31). The distance metric $d(\Sigma, \widehat{\Sigma})$ is defined based on $F(\Sigma, \widehat{\Sigma})$, $\text{KL}(\Sigma || \widehat{\Sigma})$, and $G(\Sigma, \widehat{\Sigma})$.

The optimization problems solved using MOSEK are implemented as follows. When the distance function $d(\Sigma, \widehat{\Sigma})$ is the Frobenius norm, the factor model problem (5) is directly solved via the MOSEK solver. For the case of the KL divergence and the Gelbrich distance, we solve the reformulation proposed by Lemmas A.1 and A.2, respectively. The implementations have been run on a Core(TM) i7-10610U CPU with 1.80GHz 2.30GHz clock speed, and 16GB of RAM. The SDPs are solved with MOSEK 9.3 using the YALMIP interface [43] in MATLAB R2024a. The numerical results of this experiment are illustrated in Figure 3.

According to Figure 3, although the proposed algorithm may take a higher execution time for lower-dimensional data and small ε , it is more efficient than the off-the-shelf solver MOSEK for higher-dimensional data. In more detail, our algorithm is able to solve the problem for high-dimensional settings, while MOSEK stopped solving the problem for higher-dimensional data in our hardware with the specifications mentioned above. In addition, for the cases in which $d(\Sigma, \widehat{\Sigma})$ is defined based on $F(\Sigma, \widehat{\Sigma})$ and $G(\Sigma, \widehat{\Sigma})$, for lower-dimensional data, and for big ε , the algorithm is more efficient than MOSEK, as it takes less computational time. We wish to note that in the case of the KL divergence with $\varepsilon = 10$ and dimensions larger than 150, the algorithm converges efficiently. However, it still produces a non-zero duality gap for $n = 50$ and triggers an “Out of Memory” error for $n = 100$, despite the optimal objective value being almost zero. These issues are resolved by calculating the averaged values of Σ_t to compute the objective value at each iteration for $n = 50$, and by removing the denominator in the stopping condition (31), which tends to zero for $n = 100$. We also note that for the cases in which $d(\Sigma, \widehat{\Sigma})$ is defined based on $F(\Sigma, \widehat{\Sigma})$ and $\text{KL}(\Sigma || \widehat{\Sigma})$, for the dimensions of 200 and higher, and also for the case of $d(\Sigma, \widehat{\Sigma})$ defined based on $G(\Sigma, \widehat{\Sigma})$, for the dimensions of 150 and

higher, MOSEK cannot solve the problem due to an ‘Unknown problem in solver’, corresponding to ‘YALMIP error code’ equal to 9.

5. CONCLUSION AND FUTURE DIRECTION

This paper proposes a saddle point reformulation of the factor model problem, along with a first-order algorithm to solve it. This algorithm relies on an LMO to solve the inner minimization problem in the saddle point framework. We address the inaccuracy of the empirical covariance matrix $\widehat{\Sigma}$ by robustifying the solution via a family of covariance matrices close to $\widehat{\Sigma}$ with respect to a generic distance function. To showcase the practicality of the proposed scheme, we further derive semi-closed form solutions for the LMO in three particular choices of distance functions: the Frobenius norm, the Kullback-Leibler divergence, and the Gelbrich distance.

Considering further research directions, there is a need to interpret the physical meaning of the factor model components (e.g., factor loading matrix and factors) in a dynamical system setting. Gaining this understanding can help in predicting system behavior by analyzing these components based on their physical interpretation, or by identifying their variation ranges to determine system stability or detect potential anomalies. Building on this insight, efforts could also be directed towards controller design by finding the mapping between control parameters and factor model parameters. Achieving this goal relies on addressing the interpretation challenge. Once the connection between physical behavior and factor dynamics is better understood, effective controllers can be designed to directly regulate the system through the factor model parameters. This development can ultimately lead to more robust and adaptive control strategies.

APPENDIX A. TECHNICAL LEMMAS

In this appendix, we report two technical lemmas that enable us to solve the factor model problem (5) using commercial optimization solvers such as MOSEK.

Lemma A.1 (KL-Factor model reformulation). *The factor model problem (5) with the KL divergence $d(\Sigma, \widehat{\Sigma}) = \text{KL}(\Sigma || \widehat{\Sigma})$ is equivalent to the optimization program*

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

where \mathbb{L}_n is the space of all n -dimensional lower triangular matrices.

For the proof of Lemma A.1, we refer to the MOSEK Modeling Cookbook [46, Section 6.2.3].

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

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

The proof of Lemma A.2 follows immediately from the application of [50, Proposition 2.2] to the factor model problem (5).

REFERENCES

- [1] Brian David Outram Anderson and Manfred Deistler. Identification of dynamic systems from noisy data: Single factor case. *Mathematics of Control, Signals, and Systems*, 6:10–29, 1993.
- [2] Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.
- [3] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.
- [4] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221, 2002.
- [5] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.
- [6] Dimitri Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty*. PhD thesis, Massachusetts Institute of Technology, 1971.
- [7] Dimitri Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.
- [8] Dimitri Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.
- [9] Dimitris Bertsimas, Martin S. Copenhaver, and Rahul Mazumder. Certifiably optimal low rank factor analysis. *Journal of Machine Learning Research*, 18:1–53, 2017.
- [10] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. Strong convexity of sandwiched entropies and related optimization problems. *Reviews in Mathematical Physics*, 30(09):1850014, 2018.
- [11] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [12] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.
- [13] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.
- [14] Joseph Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, 2013.
- [15] Giulio Bottegal and Giorgio Picci. Modeling complex systems by generalized factor analysis. *IEEE Transactions on Automatic Control*, 60(3):759–774, 2014.
- [16] Craig Boutilier, Thomas Dean, and Steve Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

- [17] James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference*, pages 28–47. Springer, 1986.
- [18] Eric Carlen and Elliott H Lieb. Some matrix rearrangement inequalities. *Annali di Matematica Pura ed Applicata*, 185(Suppl 5):S315–S324, 2006.
- [19] Valentina Ciccone, Augusto Ferrante, and Mattia Zorzi. An alternating minimization algorithm for factor analysis. *arXiv:1806.04433*, 2018.
- [20] Valentina Ciccone, Augusto Ferrante, and Mattia Zorzi. Factor models with real data: A robust estimation of the number of factors. *IEEE Transactions on Automatic Control*, 64(6):2412–2425, 2018.
- [21] Thomas M Cover. *Elements of Information Theory*. Wiley, 1999.
- [22] Georgios Darivianakis, Angelos Georghiou, Soroosh Shafiee, and John Lygeros. A robust optimization approach to network control using local information exchange. *Operations Research (Forthcoming)*, 2024.
- [23] Manfred Deistler, Wolfgang Scherrer, and Brian David Outram Anderson. The structure of generalized linear dynamic factor models. *Empirical Economic and Financial Research: Theory, Methods and Practice*, pages 379–400, 2015.
- [24] Ramón A. Delgado, Juan C. Agüero, and Graham C. Goodwin. A rank-constrained optimization approach: Application to factor analysis. *IFAC Proceedings Volumes*, 47(3):10373–10378, 2014.
- [25] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: B*, 39(1):1–22, 1977.
- [26] Frank Deutsch and Hein Hundal. The rate of convergence of Dykstra’s cyclic projections algorithm: The polyhedral case. *Numerical Functional Analysis and Optimization*, 15(5-6):537–565, 1994.
- [27] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
- [28] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.
- [29] Robert Engle and Mark Watson. A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76(376):774–781, 1981.
- [30] Lucia Falconi, Augusto Ferrante, and Mattia Zorzi. A robust approach to arma factor modeling. *IEEE Transactions on Automatic Control*, 69(2):828–841, 2023.
- [31] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: B*, 75(4):603–680, 2013.
- [32] Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4):540–554, 2000.
- [33] Mario Forni and Marco Lippi. The generalized dynamic factor model: Representation theory. *Econometric Theory*, 17(6):1113–1141, 2001.
- [34] Norbert Gaffke and Rudolf Mathar. A cyclic projection algorithm via duality. *Metrika*, 36(1):29–54, 1989.
- [35] Matthias Gelbrich. On a formula for the L^2 wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

- [36] Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.
- [37] Shih-Ping Han. A successive projection method. *Mathematical Programming*, 40(1):1–14, 1988.
- [38] Christiaan Heij and Wolfgang Scherrer. System identification by dynamic factor models. *SIAM Journal on Control and Optimization*, 35(6):1924–1951, 1997.
- [39] George Kapetanios and Massimiliano Marcellino. A parametric estimation method for dynamic factor models of large dimensions. *Journal of Time Series Analysis*, 30(2):208–238, 2009.
- [40] Koulik Khamaru and Rahul Mazumder. Computation of the maximum likelihood estimator in low-rank factor analysis. *Mathematical Programming*, 176:279–310, 2019.
- [41] Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.
- [42] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 535–541, 2000.
- [43] Johan Lofberg. Yalmip: A toolbox for modeling and optimization in matlab. In *International Conference on Robotics and Automation*, pages 284–289, 2004.
- [44] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46(1):157–178, 1993.
- [45] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley, 2007.
- [46] Mosek ApS. Mosek modeling cookbook: Release 3.3.1, 2025.
- [47] Angelia Nedić and Asuman Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.
- [48] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142:205–228, 2009.
- [49] Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Operations Research*, 70:490–515, 2021.
- [50] Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Bridging bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 48(1):1–37, 2023.
- [51] Giorgio Picci, Lucia Falconi, Augusto Ferrante, and Mattia Zorzi. Hidden factor estimation in dynamic generalized factor analysis models. *Automatica*, 149:110834, 2023.
- [52] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [53] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79. PMLR, 2013.
- [54] Charles Spearman. “General intelligence,” objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.
- [55] Ryan J Tibshirani. Dykstra’s algorithm, ADMM, and coordinate descent: Connections, insights, and extensions. In *Advances in Neural Information Processing Systems*, 2017.

- [56] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009.
- [57] Jan Hendrik van Schuppen. Stochastic realization problems motivated by econometric modeling. In *Modelling, Identification and Robust Control*, pages 259–275. 1986.
- [58] John von Neumann. On rings of operators. Reduction theory. *Annals of Mathematics*, 50(2):401–485, 1949.
- [59] Xiaozhou Wang and Ting Kei Pong. Convergence rate analysis of a Dykstra-type projection algorithm. *SIAM Journal on Optimization*, 34(1):563–589, 2024.
- [60] Barry M Wise, Neal B Gallagher, Stephanie Watts Butler, Daniel D White, and Gabriel G Barna. A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *Journal of Chemometrics*, 13(3-4):379–396, 1999.
- [61] Ningning Wu and Jing Zhang. Factor-analysis based anomaly detection and clustering. *Decision Support Systems*, 42(1):375–389, 2006.
- [62] Le Zhou, Yaixin Wang, Zhiqiang Ge, and Zhihuan Song. Multirate factor analysis models for fault detection in multirate processes. *IEEE Transactions on Industrial Informatics*, 15(7):4076–4085, 2018.
- [63] Mattia Zorzi and Rodolphe Sepulchre. Factor analysis of moving average processes. In *European Control Conference*, pages 3579–3584, 2015.