The Wasserstein metric in Factor Analysis*

Lipeng Ning and Tryphon Georgiou†

Abstract

We consider the problem of approximating a (nonnegative definite) covariance matrix by the sum of two structured covariances –one which is diagonal and one which has low-rank. Such an additive decomposition follows the dictum of factor analysis where linear relations are sought between variables corrupted by independent measurement noise. We use as distance the Wasserstein metric between their respective distributions (assumed Gaussian) which induces a metric between nonnegative definite matrices, in general. The rank-constraint renders the optimization non-convex. We propose alternating between optimization with respect to each of the two summands. Properties of these optimization problems and the performance of the approach are being analyzed.

Keywords

Factor analysis, optimal transport, system identification

1 Introduction

Consider a zero-mean Gaussian random vector $\mathbf{x}$ taking values in $\mathbb{R}^{n\times 1}$ and with covariance Σ. We assume that $\mathbf{x}$ is consistent with the basic assumptions of factor analysis and of errors-in-variables models (see e.g., [1, 5, 7]), namely that

$$\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}}$$

where $\hat{\mathbf{x}}, \tilde{\mathbf{x}}$ are both zero-mean independent (Gaussian) random vectors in $\mathbb{R}^{n\times 1}$ with

$$\begin{aligned} &\mathcal{E}(\hat{\mathbf{x}}\hat{\mathbf{x}}) = \hat{\Sigma},\ \mathrm{rank}(\hat{\Sigma}) = r < n \\ &\mathcal{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}') = D \text{ is diagonal} \\ &\mathcal{E}(\hat{\mathbf{x}}\tilde{\mathbf{x}}') = 0, \end{aligned}$$

where $\mathcal{E}$ denotes expectation. The entries of $\tilde{\mathbf{x}}$ represent independent "measurement noise", while the rank deficiency of the covariance of the "noise-free" component $\hat{\mathbf{x}}$ indicates that linear relations are satisfied by its entries. Factor analysis aims at identifying consistent decompositions of Σ into $\hat{\Sigma} + D$ with the above properties and with r small, so as to reliably identify linear relations from observational data and empirical statistics.

Since the covariance Σ is often approximated by

$$\frac{1}{N}\sum_{k=1}^{N} x_k x_k',$$

where $x_1, \ldots, x_N$ represent independent vector-measurements of $\mathbf{x}$, it is common to estimate $\hat{\Sigma}$ and D via optimization of a likelihood function or, via minimization of some "distance" between Σ and the sum $\hat{\Sigma} + D$ of the required form (see, e.g., [4, 3]). It should be noted that logarithmic distances (Kullback-Leibler, Thompson) require that Σ is invertible. However, when the number N of available samples is small (smaller than n) this requirement is not satisfied. Herein, we explore an alternative distance between covariance matrices which is induced by the Wasserstein distance (Section 2) of their corresponding probability distributions and is not limited in this respect. This provides a metric between covariances (Section 2) which is amenable to tools from convex analysis when seeking decompositions of a sample covariance in accordance to the factor analysis model (Sections 3 and 4).

2 Optimal transport

The Monge-Kantorovich transportation problem seeks an optimal transference plan for moving a given mass/probability distribution $p_{\mathbf{x}}$ to another distribution $p_{\mathbf{y}}$ (see [11]). These can be thought as marginals of two jointly distributed random variables $\mathbf{x}, \mathbf{y}$, in which case the transportation cost can be written as $\mathcal{E}\left(\mathrm{cost}(\mathbf{x}-\mathbf{y})\right)$. The transference plan relates to a choice of a compatible joint distribution, having the given marginals, which minimizes the cost. We will only be concerned with a quadratic cost $\mathcal{E}\left(\|\mathbf{x}-\mathbf{y}\|^2\right)$ which induces the so-called Wasserstein distance between the two marginals, namely

$$\begin{aligned} W_2(p_{\mathbf{x}}, p_{\mathbf{y}}) \quad := \quad & \inf_{p(x,y)} \left\{ \sqrt{\mathcal{E}\left(\|\mathbf{x}-\mathbf{y}\|^2\right)} \mid \right. \\ & \left. \int_x p(x,y)dx = p_{\mathbf{y}}(y),\ \int_y p(x,y)dy = p_{\mathbf{x}}(x) \right\} \end{aligned}$$

*Supported by the AFOSR (FA9550-12-1-0319), the NSF (Grant 1027696), and the Vincentine Hermes-Luh Endowment.

†Dept. of Electrical & Comp. Eng., University of Minnesota, Minneapolis, MN 55455; {ningx015, tryphon}@umn.edu.

where $p(x, y)$ is a joint distribution (possibly, measure).

If $\mathbf{x}$, $\mathbf{y}$ are jointly distributed zero-mean and Gaussian with covariances $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$, respectively, and cross-covariance $C_{xy} = \mathcal{E}(\mathbf{x}\mathbf{y}')$, then

$$\mathcal{E}(\|\mathbf{x} - \mathbf{y}\|^2) = \mathrm{tr}(\Sigma_x + \Sigma_y - C_{xy} - C'_{xy}).$$

Hence, the Wasserstein distance can be obtained by solving the following optimization problem:

$$\min_C \left\{ \sqrt{\mathrm{tr}(\Sigma_x + \Sigma_y - 2C)} \mid \begin{bmatrix} \Sigma_x & C \\ C' & \Sigma_y \end{bmatrix} \geq 0 \right\}. \tag{2.1}$$

The minimal value of (2.1) can be readily expressed as

$$\sqrt{\mathrm{tr}(\Sigma_x + \Sigma_y - 2(\Sigma_y^{\frac{1}{2}}\Sigma_x\Sigma_y^{\frac{1}{2}})^{\frac{1}{2}})} =: d(\Sigma_x, \Sigma_y) \tag{2.2}$$

(see [9, 6, 8]). This defines a metric between covariances that are not restricted to be invertible (as is the case with e.g., the Thompson metric).

3 Factor analysis

We now consider the additive decomposition of $\Sigma \geq 0$ into a sum of two covariances as postulated by Factor Analysis –one having low-rank and another being diagonal, both of the same size as Σ:

$$\min_{\hat{\Sigma}, D} \left\{ d(\Sigma, \hat{\Sigma} + D) \mid D \geq 0 \text{ diagonal}, \; \hat{\Sigma} \geq 0, \mathrm{rank}(\hat{\Sigma}) = r \right\} \tag{3.3a}$$

$$= \min_{\hat{\Sigma}, D, C} \left\{ \sqrt{\mathrm{tr}(\Sigma + \hat{\Sigma} + D - 2C)} \mid \begin{bmatrix} \Sigma & C \\ C' & \hat{\Sigma} + D \end{bmatrix} \geq 0, \; D \geq 0, \; D \text{ diagonal}, \hat{\Sigma} \geq 0, \; \mathrm{rank}(\hat{\Sigma}) = r \right\}. \tag{3.3b}$$

The square of the objective function is linear and all constraints except for the rank are convex. Interestingly, when either of $D, \hat{\Sigma}$ is set to zero, the above minimization problem can be solved efficiently. We consider each of these two cases separately first.

3.1 Approximation with a diagonal matrix

Starting with $\Sigma \geq 0$, we formulate the problem

$$\begin{aligned} &\min\{d(\Sigma, D) \mid D \geq 0, \; D \text{ diagonal}\} \\ &= \min_{D, C} \left\{ \mathrm{tr}(\Sigma + D - C - C') \mid D \text{ diagonal}, \; \begin{bmatrix} \Sigma & C \\ C' & D \end{bmatrix} \geq 0 \right\}. \end{aligned} \tag{3.4}$$

This is convex and can be solved efficiently [2]. Next, we analyze the structure of the minimizers.

PROPOSITION 1. *Given $\Sigma \geq 0$, let D_{opt} be a minimizer of (3.4). Then, there exists a matrix $\Lambda_{\mathrm{opt}} \geq 0$ with $[\Lambda_{\mathrm{opt}}]_{ii} = 1$, for $i = 1, \ldots, n$, such that*

$$\Sigma = \Lambda_{\mathrm{opt}} D_{\mathrm{opt}} \Lambda_{\mathrm{opt}}. \tag{3.5}$$

When $\Sigma > 0$, D_{opt} and Λ_{opt} are uniquely defined.

Note that any covariance Σ can be expressed as $\Sigma = D_1 \Lambda_1 D_1$, with D_1 diagonal and Λ_1 a correlation matrix, i.e., a covariance with ones on the diagonal. Observe that a complementary factorization also exists (3.5) and relates to the solution of (3.4).

Proof. [Proof of Proposition 1:] Let

$$\Sigma = \begin{bmatrix} U & V \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U' \\ V' \end{bmatrix}$$

be the eigen-decompositon of Σ and $E > 0$. Then

$$\begin{bmatrix} \Sigma & C \\ C' & D \end{bmatrix} \geq 0 \;\Rightarrow\; \left[\begin{array}{cc:c} E & 0 & U'C \\ 0 & 0 & V'C \\ \hdashline C'U & C'V & D \end{array}\right] \geq 0.$$

Then, $V'C = 0$ while (3.4) becomes

$$\min_{D, C} \left\{ \mathrm{tr}(D - C - C') + \mathrm{tr}(E) \mid V'C = 0, \; \begin{bmatrix} E & U'C \\ C'U & D \end{bmatrix} \geq 0, D \text{ diagonal} \right\}. \tag{3.6}$$

Let $\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda'_{12} & \Lambda_{22} \end{bmatrix} \geq 0$ and W be multipliers for the first two constraints of (3.6) with Λ_{11} of the same size as E. The corresponding Lagrangian is

$$\begin{aligned} \mathcal{L} &= \mathrm{tr}(D - C - C') + \mathrm{tr}(E) + \mathrm{tr}(WV'C) \\ &\quad - \mathrm{tr}\left(\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda'_{12} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} E & U'C \\ C'U & D \end{bmatrix} \right) \\ &= \mathrm{tr}\left(2\left((-I - [\Lambda'_{12} \; W/2] \begin{bmatrix} U' \\ V' \end{bmatrix}) C \right) + (I - \Lambda_{22}) D \right) \\ &\quad + \mathrm{tr}((I - \Lambda_{11})E). \end{aligned}$$

Then

$$\begin{aligned} &[\Lambda'_{12} \; W/2] \begin{bmatrix} U' \\ V' \end{bmatrix} = -I \;\Rightarrow\; \Lambda_{12} = -U' \\ &[\Lambda_{22}]_{ii} = 1, \; \forall i = 1, \ldots, n. \end{aligned}$$

The dual problem to (3.6) is

$$\max_{\Lambda_{11}, \Lambda_{22}} \left\{ \mathrm{tr}\left((I - \Lambda_{11})E\right) \mid \begin{bmatrix} \Lambda_{11} & -U' \\ -U & \Lambda_{22} \end{bmatrix} \geq 0, \; [\Lambda_{22}]_{ii} \leq 1, \forall i = 1, \ldots, n \right\}. \tag{3.7}$$

Clearly, the optimal value is obtained when $\Lambda_{11} = U'\Lambda_{22}^{\dagger}U$ where $\Lambda_{22}^{\dagger}$ denotes the Moore-Penrose pseudo inverse of Λ_{22} which coincides with the inverse when Λ_{22} is invertible. Let $\Lambda_{22} = \Lambda_{\rm opt}$ be a minimizer of (3.7). Since (3.6) is strictly feasible, the duality gap between (3.7) and (3.6) is zero. The following condition is satisfied

$$\begin{bmatrix} U'\Lambda_{\rm opt}^{\dagger}U & -U' \\ -U & \Lambda_{\rm opt} \end{bmatrix} \begin{bmatrix} E & U'C_{\rm opt} \\ C'_{\rm opt}U & D_{\rm opt} \end{bmatrix} = 0,$$

where $(D_{\rm opt}, C_{\rm opt})$ is a minimizer of (3.6). Then,

$$\Lambda_{\rm opt}C'_{\rm opt}U = UE \;\Rightarrow\; \Lambda_{\rm opt}C'_{\rm opt}UU' = \Sigma$$
$$\Lambda_{\rm opt}D_{\rm opt} = UU'C_{\rm opt}.$$

Thus, (3.5) holds.

If Σ is invertible, then U is an orthogonal matrix and Λ_{22} in (3.6) is also invertible. By substituting $\Lambda_{11} = U'\Lambda_{22}^{-1}U$ into (3.6), we see that the optimal value of Λ_{22} is the minimizer of the following problem

$$\min_{\Lambda_{22}} \left\{ \mathrm{tr}(\Lambda_{22}^{-1}\Sigma) \mid \Lambda_{22} \geq 0, [\Lambda_{22}]_{ii} = 1, \forall i = 1, \ldots, n \right\}.$$

The latter is strictly convex and therefore it has a unique minimizer. □

We note that (3.4) has a solution even when Σ is singular. However, in this case, the solution may not be unique. To see this, take $\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Then, any diagonal $D \geq 0$ with $\mathrm{tr}(D) = 1$ is a minimizer of (3.4).

3.2 Approximation with a low-rank matrix

Given $\Sigma \geq 0$, as noted earlier, the problem

$$\begin{aligned} (3.8)\quad & \min\{d(\Sigma, \hat{\Sigma}) \mid \hat{\Sigma} \geq 0 \text{ with } \mathrm{rank}(\hat{\Sigma}) = r\} \\ &= \min_{\hat{\Sigma}, C} \left\{ \mathrm{tr}(\Sigma + \hat{\Sigma} - C - C') \mid \begin{bmatrix} \Sigma & C \\ C' & \hat{\Sigma} \end{bmatrix} \geq 0, \right. \\ &\qquad\qquad \left. \mathrm{rank}(\hat{\Sigma}) = r \right\} \end{aligned}$$

is not convex. Yet, as we show below, a solution can be readily obtained from the spectral decomposition of Σ.

LEMMA 3.1. *Let Σ, $\hat{\Sigma} \geq 0$ and of equal size, and let $\hat{\Pi}$ denote the orthogonal projection onto the range of $\hat{\Sigma}$. The following holds:*

$$(3.9)\qquad d(\Sigma, \hat{\Sigma}) = d(\hat{\Pi}\Sigma\hat{\Pi}, \hat{\Sigma}) + \mathrm{tr}\left((I - \hat{\Pi})\Sigma\right).$$

Proof. Clearly,

$$\begin{aligned} d(\Sigma, \hat{\Sigma}) &= \mathrm{tr}(\Sigma + \hat{\Sigma} - 2(\hat{\Sigma}^{\frac{1}{2}}\Sigma\hat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &= \mathrm{tr}(\hat{\Pi}\Sigma\hat{\Pi} + \hat{\Sigma} - 2(\hat{\Sigma}^{\frac{1}{2}}\hat{\Pi}\Sigma\hat{\Pi}\hat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &\quad + \mathrm{tr}((I - \hat{\Pi})\Sigma) \\ &= d(\hat{\Pi}\Sigma\hat{\Pi}, \hat{\Sigma}) + \mathrm{tr}\left((I - \hat{\Pi})\Sigma\right). \end{aligned}$$

□

PROPOSITION 2. *Let $\Sigma \geq 0$ with eigenvalues $\sigma_1 \geq \ldots \geq \sigma_n$ and corresponding eigenvectors $v_1, \ldots v_n$. Then,*

$$\min\{d(\Sigma, \hat{\Sigma}) \mid \hat{\Sigma} \geq 0 \text{ with } \mathrm{rank}(\hat{\Sigma}) = r\} = \sum_{k=r+1}^{n} \sigma_k$$

and the minimum is attained for $\hat{\Sigma} = \sum_{k=1}^{r} \sigma_k v_k v_k'$.

Proof. Let $\hat{\Pi}$ denote the orthogonal projection to the range of the optimizer $\hat{\Sigma}_{\rm opt}$. From Lemma 3.1,

$$\begin{aligned} d(\Sigma, \hat{\Sigma}_{\rm opt}) &= d(\hat{\Pi}\Sigma\hat{\Pi}, \hat{\Sigma}_{\rm opt}) + \mathrm{tr}\left((I - \hat{\Pi})\Sigma\right) \\ &\geq \mathrm{tr}\left((I - \hat{\Pi})\Sigma\right) = \sum_{k=r+1}^{n} \sigma_k. \end{aligned}$$

The above holds with $\hat{\Sigma}_{\rm opt}$ as in the statement. □

4 Stationarity conditions

Next we consider the more general problem (3.3). Since any feasible $\hat{\Sigma}$ can be expressed as

$$\hat{\Sigma} = FF' \text{ with } F \in \mathbb{R}^{n \times r},$$

we can replace (3.3) by

$$(4.10)\qquad \min_{F,D} \mathrm{tr}\left(\Sigma + FF' + D - 2(\Sigma^{\frac{1}{2}}(FF' + D)\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\right).$$

Using variational analysis and assuming that $FF' + D$ and Σ are invertible, we obtain the following necessary conditions for FF' and D to correspond to a minimizer of (4.10):

$$(4.11a)\qquad \left(\Sigma^{\frac{1}{2}}\left(\Sigma^{\frac{1}{2}}(FF' + D)\Sigma^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\right)F = F$$

$$(4.11b)\qquad [\Sigma^{\frac{1}{2}}\left(\Sigma^{\frac{1}{2}}(FF' + D)\Sigma^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}]_{ii} \begin{cases} = 1 & \text{if } [D]_{ii} > 0 \\ \leq 1 & \text{if } [D]_{ii} = 0. \end{cases}$$

Unfortunately, the objective function in (4.10) is not convex in F and D. (To see this, specialize to the case where D and F are scalars.) Due to this, no general approach is available for solving (4.11) directly. In the following we reformulate the problem so as to take advantage of the special cases in Sections 3.2 and 3.1, and obtain candidate solutions by alternating descent.

5 Reformulation and alternating descent

We begin with a reformulation of Problem (3.3b).

LEMMA 5.1. *Let $\Sigma \geq 0$ and $(\hat{\Sigma}_{\rm opt}, D_{\rm opt})$ be a minimizer*

of (3.3b). *Then* $(\hat{\Sigma}_{\rm opt}, D_{\rm opt})$ *is also a minimizer of*

$$\min_{\hat{\Sigma},D,C_1,C_2} \Big\{ \mathrm{tr}(\Sigma+\hat{\Sigma}+D-2(C_1+C_2)) \mid \begin{bmatrix} \Sigma & C_1 & C_2 \\ C_1' & D & 0 \\ C_2' & 0 & \hat{\Sigma} \end{bmatrix} \geq 0,\ D \geq 0, \\ D \text{ diagonal}, \hat{\Sigma} \geq 0,\ \mathrm{rank}(\hat{\Sigma}) = r \Big\}. \tag{5.12}$$

Proof. For $C = C_1 + C_2$,

$$\begin{bmatrix} \Sigma & C_1 & C_2 \\ C_1' & D & 0 \\ C_2' & 0 & \hat{\Sigma} \end{bmatrix} \geq 0 \Rightarrow \begin{bmatrix} \Sigma & C \\ C' & D+\hat{\Sigma} \end{bmatrix} \geq 0.$$

Hence, the optimal value of (3.3b) is no larger than that of (5.12). Now, for any triple $(\hat{\Sigma}, D, C)$ that satisfies

$$\begin{bmatrix} \Sigma & C \\ C' & D+\hat{\Sigma} \end{bmatrix} \geq 0,$$

let $C_1 = C(D+\hat{\Sigma})^{-1}D$ and $C_2 = C(D+\hat{\Sigma})^{-1}\hat{\Sigma}$. Then

$$\begin{bmatrix} \Sigma & C_1 & C_2 \\ C_1' & D & 0 \\ C_2' & 0 & \hat{\Sigma} \end{bmatrix} \geq 0.$$

Thus, if $(\hat{\Sigma}_{\rm opt}, D_{\rm opt}, C_{\rm opt})$ minimizes (3.3b), then $(\hat{\Sigma}_{\rm opt}, D_{\rm opt})$ minimizes (5.12) with corresponding C_1 and C_2 computed as above. The minimal values in (3.3b) and (5.12) are identical. □

In view of the above, we observe the following. If $(\hat{\Sigma}_{\rm opt}, D_{\rm opt})$, $C_{1,\rm opt}$ and $C_{2,\rm opt}$ are minimizers of (5.12), we can fix the value for $\hat{\Sigma}$ to this optimal value $\hat{\Sigma}_{\rm opt}$ and then $D_{\rm opt}$, $C_{1,\rm opt}$ and $C_{2,\rm opt}$ can be recovered by solving

$$\min_{D,C_1,C_2} \Big\{ \mathrm{tr}(\Sigma+\hat{\Sigma}_{\rm opt}+D-2(C_1+C_2)) \mid \begin{bmatrix} \Sigma & C_1 & C_2 \\ C_1' & D & 0 \\ C_2' & 0 & \hat{\Sigma}_{\rm opt} \end{bmatrix} \geq 0,\ D \geq 0 \text{ diagonal} \Big\}. \tag{5.13}$$

Alternatively, since

$$\begin{bmatrix} \Sigma & C_1 & C_2 \\ C_1' & D & 0 \\ C_2' & 0 & \hat{\Sigma} \end{bmatrix} \geq 0 \Rightarrow \begin{bmatrix} D & 0 & 0 \\ 0 & \Sigma - C_1 D^{-1} C_1' & C_2 \\ 0 & C_2' & \hat{\Sigma} \end{bmatrix} \geq 0,$$

if we fix $D = D_{\rm opt}$ and $C_1 = C_{1,\rm opt}$, we can recover $\hat{\Sigma}_{\rm opt}$ from $\Sigma - C_{1,\rm opt} D_{\rm opt}^{-1} C_{1,\rm opt}'$ via truncated eigenvalue decomposition (see Proposition 2).

Thus, we seek candidate solutions for (3.3b) and (5.12) by alternating between (3.4) and (3.8) for D and $\hat{\Sigma}$, respectively. In summary, we start from an initial pair $\hat{\Sigma}_{(0)}$ and $D_{(0)}$. In the k-th iteration for $k \geq 1$:

i) obtain $D_{(k)}$, $C_{1,(k)}$ and $C_{2,(k)}$ via (5.13) with $\hat{\Sigma}_{\rm opt}$ replaced by $\hat{\Sigma}_{(k-1)}$,

ii) obtain $\hat{\Sigma}_{(k)}$ by truncating the eigenvalue decomposition of $\Sigma - C_{1,(k)} D_{(k)}^{-1} C_{1,(k)}'$.

Since the objective function is reduced in each iteration and the function is bounded from below by zero, the algorithm converges (but not necessarily to a globally optimal point).

Example 1: We highlight the algorithm on an academic example. We take $\Sigma > 0$ of size 50×50 in the form

$$\Sigma = FF' + D,$$

where F is 50×5 and D is a diagonal matrix. The entries of F are taken as independent samples from the standard Gaussian distribution (mean zero and variance one) while the diagonal entries of D are independently and uniformly sampled from the interval [0, 10]. The seeds in Matlab that we used to generate this example for F and D are 100 and 123, respectively. We compute $\hat{\Sigma}_{(0)}$ by truncating small eigenvalues of Σ. The initial $D_{(0)}$ is set to be the zero matrix. Figure 1 displayes the values $d(\Sigma, \hat{\Sigma}_{(k)} + D_{(k)})$ in each iterations on a logarithmic scale.

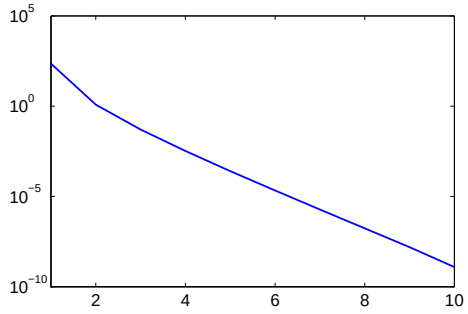

Figure 1: $\log\left(d(\Sigma, \hat{\Sigma}_{(k)} + D_{(k)})\right)$ vs. k

Example 2: We conclude with an academic factor-analysis problem. Consider

$$X = FU + V,$$

where $F \in \mathbb{R}^{20\times 5}$, $U \in \mathbb{R}^{5\times 50}$ and $V \in \mathbb{R}^{20\times 50}$. The entries of F and U are taken as independent samples from a standard Gaussian distribution, and each column of the noise-component V is taken from a multivariable Gaussian distribution with zero-mean

and diagonal covariance D. The entries $[D]_{ii}$ of the noise covariance are themselves sampled from a uniform distribution on $[\sigma, 2\sigma]$. Then, the diagonal entries of $FUU'F'/50$ have mean 5 while the diagonal entries of $VV'/50$ have mean $3\sigma/2$. Thus, the signal-to-noise ratio (SNR) is approximately $10/3\sigma$. We thus obtain a sample covariance matrix and re-scale it to a correlation matrix Σ.

Next, we approximate Σ as a sum of a singular covariance with rank 5 and a diagonal "noise"-covariance following our approach and we compare with two other standard techniques. In particular, we compare with a maximum-likelihood-based method proposed in [4]. This is the basis of the Matlab routine *factoran*. We also compare with the "total-least-squares" which is based on the eigen-decomposition of Σ and retention of the corresponding "dominant" subspace. One should note that in this last method the structure of the noise covariance (diagonal) is not taken into account at all.

We denote by $\hat{\Sigma}_{\text{Tran}}$, $\hat{\Sigma}_{\text{ML}}$ and $\hat{\Sigma}_{\text{TLS}}$ the estimated low-rank covariance matrices in these three methods, respectively. We assess the performance of each using the gap distance between the true and estimated covariance matrices (this is the angle between their range spaces). To this end, let Π_{true} denote the orthogonal projection onto the range of F, and let Π_{Tran}, Π_{ML} and Π_{TLS} denote the projection onto the range of the respective low-rank approximations of XX'. The gap metric between corresponding range spaces is $\|\Pi_{\text{estimate}} - \Pi_{\text{true}}\|$. where Π_{estimate} represents the projection onto the range space of an estimated low-rank approximate covariance. The gap represents the sign of the principle angle between the two subspaces (see, [10, page 93]). We choose a range of values for σ between 1/2 and 10 so that the SNR is between 1/3 to 20/3. In each instance we run 100 simulations, compute the corresponding low-rank matrices and evaluate the gap distances to the subspace corresponding to the scaled F ("true"). The average gap distance in these 100 simulations is tabulated in Figure 2 for each of the three methods. We observe that the optimal-transport-based approach and the maximum-likelihood-based one have similar performance and that they are better than the total-least-squares method, especially in high SNR cases. In general, the transportation-based approach appears to have a slight advantage over the maximum-likelihood method.

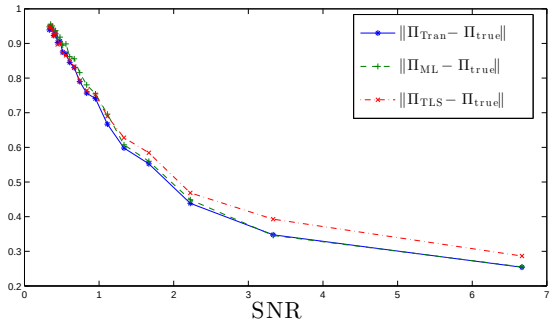

Figure 2: $\|\Pi_{\text{Tran}} - \Pi_{\text{true}}\|$, $\|\Pi_{\text{ML}} - \Pi_{\text{true}}\|$ and $\|\Pi_{\text{TLS}} - \Pi_{\text{true}}\|$ vs. SNR

References

[1] T. Anderson and H. Rubin, *Statistical inference in factor analysis*, in Proceedings of the third Berkeley symposium on mathematical statistics and probability, vol. 5, 1956, pp. 111–150.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[3] H. Harman and W. Jones, *Factor analysis by minimizing residuals (minres)*, Psychometrika, 31 (1966), pp. 351–368.

[4] K. Jöreskog, *Some contributions to maximum likelihood factor analysis*, Psychometrika, 32 (1967), pp. 443–482.

[5] S. Klepper and E. Leamer, *Consistent sets of estimates for regressions with errors in all variables*, Econometrica: Journal of the Econometric Society, (1984), pp. 163–183.

[6] M. Knott and C. S. Smith, *On the optimal mapping of distributions*, Journal of Optimization Theory and Applications, 43 (1984), pp. 39–49.

[7] L. Ning, T. T. Georgiou, A. Tannenbaum, and S. P. Boyd, *Linear models based on noisy data and the Frisch scheme*, Arxiv preprint, (2013).

[8] L. Ning, X. Jiang, and T. Georgiou, *Geometric methods for estimation of structured covariances*, Arxiv preprint arXiv:1110.3695, (2011).

[9] I. Olkin and F. Pukelsheim, *The distance between two random vectors with given dispersion matrices*, Linear Algebra and its Appl., 48 (1982), pp. 257–263.

[10] G. W. Stewart and J.-G. Sun, *Matrix perturbation theory*, Academic press, 1990.

[11] C. Villani, *Topics in optimal transportation*, vol. 58, American Mathematical Society, 2003.