Factor Analysis and Alternating Minimization

Lorenzo Finesso[1] and Peter Spreij[2]

[1] Institute of Biomedical Engineering, CNR-ISIB, Padova, Italy
lorenzo.finesso@isib.cnr.it
[2] Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, Amsterdam, The Netherlands
spreij@science.uva.nl

Dedicated to Giorgio Picci on the occasion of his 65th birthday.
Happy Birthday Giorgio!

1 Introduction

Factor analysis, in its original formulation, deals with the linear statistical model

$$Y = HX + \varepsilon \tag{1}$$

where H is a deterministic matrix, X and ε independent random vectors, the first with dimension smaller than Y, the second with independent components. What makes this model attractive in applied research is the *data reduction* mechanism built in it. A large number of observed variables Y are explained in terms of a small number of unobserved (latent) variables X perturbed by the independent noise ε. Under normality assumptions, which are the rule in the standard theory, all the laws of the model are specified by covariance matrices. More precisely, assume that X and ε are zero mean independent normal vectors with $\mathbb{C}\mathrm{ov}(X) = P$ and $\mathbb{C}\mathrm{ov}(\varepsilon) = D$, where D is diagonal. It follows from (1) that $\mathbb{C}\mathrm{ov}(Y) = HPH^\top + D$.

Building a factor analysis model of the observed data requires the solution of a difficult algebraic problem. Given Σ_0, the covariance matrix of Y, find the triples (H, P, D) such that $\Sigma_0 = HPH^\top + D$. Due to the structural constraint on D, which is assumed to be diagonal, the existence and unicity of a factor analysis model are not guaranteed. As it turns out, the right tools to deal with this situation come from the theory of stochastic realization, see [5] (trying to spot the master's hand) for an early contribution on the subject.

In the present paper we make a first attempt at understanding how to build an optimal *approximate* factor analysis model. The criterion we have chosen to evaluate the distance between covariances is the I-divergence between the corresponding normal laws. The algorithm that we propose for the construction of the best approximation is inspired by the alternating minimization procedure of [4] and [6].

A. Chiuso et al. (Eds.): Modeling, Estimation and Control, LNCIS 364, pp. 85–96, 2007.
springerlink.com

2 The Model

Consider two independent, zero mean, normal vectors X and ε of respective dimensions k and n. We will assume that $\mathbb{C}\mathrm{ov}(X) = I$, the identity matrix, and $\mathbb{C}\mathrm{ov}(\varepsilon) = D > 0$, a diagonal matrix. Let H be an $n \times k$ matrix (in this paper $k < n$) and let the random vector Y be defined by

$$Y = HX + \varepsilon. \tag{2}$$

Under these assumptions (2) is called a factor analysis (FA) model of size k for the vector Y. Notice that allowing $\mathbb{C}\mathrm{ov}(X) = P > 0$ does not produce a more general model, as a square root of P can always be absorbed in H. We will say that a normal vector Y admits a FA model of size k if it is equal in distribution to $HX + \varepsilon$ for some X and ε as above, i.e. if its covariance Σ_0 can be written as $\Sigma_0 = HH^\top + D$. Not every normal vector Y admits a FA model, the hard constraint being imposed by the diagonal structure of D. A probabilistic interpretation stems from $\mathbb{C}\mathrm{ov}(Y|X) = D$ (see equation (28) of the Appendix) i.e. the n components of Y are conditionally independent given the $k < n$ components of some vector X. In Remark 1 of the next section the condition for the existence of a FA model is slightly reformulated.

Although the construction of an exact FA model is not always possible, one can search for a best approximate model, according to some criterion. In this paper we opt for minimizing the I-divergence (Kullback-Leibler distance) between normal laws. Recall that given two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$, defined on the same measurable space, such that $\mathbb{P}_1 \ll \mathbb{P}_2$, the I-divergence of $\mathbb{P}_1$ with respect to $\mathbb{P}_2$ is defined as

$$D(\mathbb{P}_1||\mathbb{P}_2) = \mathbb{E}_{\mathbb{P}_1} \log \frac{d\mathbb{P}_1}{d\mathbb{P}_2}.$$

If $\mathbb{P}_1$ and $\mathbb{P}_2$ are normal measures on the same space $\mathbb{R}^n$, with zero means and strictly positive covariance matrices Σ_1 and Σ_2 respectively, the I-divergence $D(\mathbb{P}_1||\mathbb{P}_2)$ takes the explicit form

$$D(\mathbb{P}_1||\mathbb{P}_2) = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \frac{n}{2}. \tag{3}$$

Since the I-divergence only depends on the covariance matrices, we usually write $D(\Sigma_1||\Sigma_2)$ instead of $D(\mathbb{P}_1||\mathbb{P}_2)$.

The approximate factor analysis problem can be posed as follows:

Problem 1. Given the positive covariance matrix $\Sigma_0 \in \mathbb{R}^{n\times n}$ and the integer $k < n$ minimize

$$D(\Sigma_0||HH^\top + D) = \frac{1}{2} \log \frac{|HH^\top + D|}{|\Sigma_0|} + \frac{1}{2} \mathrm{tr}((HH^\top + D)^{-1}\Sigma_0) - \frac{n}{2}.$$

over all pairs (H, D) where $H \in \mathbb{R}^{n\times k}$ and $D > 0$ is of size n and diagonal.

Notice that $D(\Sigma_1||\Sigma_2)$, computed as in (3), can be considered as a divergence between two positive definite matrices, without referring to normal distributions. Hence Problem 1 also has a meaning, when one refrains from assumptions like normality.

Existence of the minimum is guaranteed by the following

Proposition 1. *There exist matrices $H^* \in \mathbb{R}^{n\times k}$ and $D^* > 0$ of size n and diagonal minimizing the I-divergence in Problem 1.*

The proof is deferred to section 4.2, since it uses later results.

In order to construct an algorithm for the solution of Problem 1 we will imitate the approach of [6]. The algorithm will therefore be derived by a relaxation technique, lifting the original minimization problem to a higher dimensional space. In the larger space a double minimization problem equivalent to Problem 1 can be formulated, leading in a natural way to an alternating minimization algorithm.

3 Lifting of the Original Problem

In this section we will embed Problem 1 into a higher dimensional space. First we introduce the relevant sets of covariances. Given $k < n$ we denote by

$$\boldsymbol{\Sigma} = \{\Sigma \in \mathbb{R}^{(n+k)\times(n+k)} : \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} > 0\}. \tag{4}$$

where Σ_{11} is $n \times n$. Two subsets of $\boldsymbol{\Sigma}$ will play a special role.

$$\boldsymbol{\Sigma}_0 = \{\Sigma \in \boldsymbol{\Sigma} : \Sigma_{11} = \Sigma_0\}. \tag{5}$$

where Σ_0 is a given covariance. We also consider the subset

$$\boldsymbol{\Sigma}_1 = \{\Sigma \in \boldsymbol{\Sigma} : \Sigma = \begin{pmatrix} HH^\top + D & HQ \\ (HQ)^\top & Q^\top Q \end{pmatrix}\}, \tag{6}$$

where $H \in \mathbb{R}^{n\times k}, Q \in \mathbb{R}^{k\times k}$ invertible, $D > 0$ diagonal. Elements of $\boldsymbol{\Sigma}_1$ will often be denoted by $\Sigma(H, D, Q)$.

Remark 1. Notice that a normal vector Y, with $\mathbb{C}\mathrm{ov}(Y) = \Sigma_0$, admits a FA model of size k iff $\boldsymbol{\Sigma}_0 \cap \boldsymbol{\Sigma}_1 \neq \emptyset$. Supposing that this is the case, take $\Sigma \in \boldsymbol{\Sigma}_0 \cap \boldsymbol{\Sigma}_1$ then, for some (H, D, Q), one has

$$\Sigma = \begin{pmatrix} \Sigma_0 & HQ \\ (HQ)^\top & Q^\top Q \end{pmatrix} = \begin{pmatrix} HH^\top + D & HQ \\ (HQ)^\top & Q^\top Q \end{pmatrix} > 0.$$

is a *bonafide* covariance of a normal vector V of dimension $n + k$. Partition $V^\top = (Y^\top, Z^\top)^\top$. It is easy to verify that $\mathbb{C}\mathrm{ov}(Y) = \Sigma_0 = HH^\top + D$ is the same as $\mathbb{C}\mathrm{ov}(HX + \varepsilon)$ for some X standard normal and ε normal, independent from X, and with diagonal covariance D.

The lifted minimization problem can be posed as follows

Problem 2

$$\min_{\Sigma' \in \boldsymbol{\Sigma}_0, \Sigma_1 \in \boldsymbol{\Sigma}_1} D(\Sigma' || \Sigma_1)$$

which can be viewed as an iterated minimization problem over each of the variables. The two resulting partial minimization problems will be investigated in the following sections. In section 3.3 we will show the connection between Problems 1 and 2. More precisely, we will prove

Proposition 2. *Let Σ_0 be given. It holds that*

$$\min_{H,D} D(\Sigma_0||HH^\top + D) = \min_{\Sigma' \in \boldsymbol{\Sigma}_0, \Sigma_1 \in \boldsymbol{\Sigma}_1} D(\Sigma'|\Sigma_1).$$

3.1 The First Partial Minimization Problem

In this section we consider the first of the two partial minimization problems. Here we minimize, for a given positive definite matrix $\Sigma \in \boldsymbol{\Sigma}$, the divergence $D(\Sigma'||\Sigma)$ over $\Sigma' \in \boldsymbol{\Sigma}_0$. The unique solution to this problem can be computed analytically and follows from

Lemma 1. *Let (Y, X) be a random vector distributed according to some $Q = Q^{Y,X}$ and let $\boldsymbol{\mathcal{P}}$ the set of all distributions $P = P^{Y,X}$ whose marginal $P^Y = P_0$, for some fixed $P_0 \ll Q^Y$. Then $\min_{P \in \boldsymbol{\mathcal{P}}} D(P||Q) = D(P^*||Q)$ where P^* is given by the Radon-Nikodym derivative*

$$\frac{dP^*}{dQ} = \frac{dP_0}{dQ^Y}.$$

Moreover,

$$D(P^*||Q) = D(P_0||Q^Y). \tag{7}$$

and, for any other $P \in \boldsymbol{\mathcal{P}}$, one has the Pythagorean law

$$D(P||Q) = D(P||P^*) + D(P^*||Q). \tag{8}$$

Proof. First we show that (7) holds. Recall that Y has law P_0 under P^*, then

$$D(P^*||Q) = \mathbb{E}_{P^*} \log \frac{dP^*}{dQ} = \mathbb{E}_{P^*} \log \frac{dP_0}{dQ^Y} = \mathbb{E}_{P_0} \log \frac{dP_0}{dQ^Y} = D(P_0||Q^Y).$$

To show that P^* is a minimizer it is clearly sufficient to prove that (8) holds.

$$\begin{aligned} D(P||Q) &= \mathbb{E}_P \log \frac{dP}{dP^*} + \mathbb{E}_P \log \frac{dP^*}{dQ} \\ &= D(P||P^*) + \mathbb{E}_P \log \frac{dP_0}{dQ^Y} \\ &= D(P||P^*) + \mathbb{E}_{P_0} \log \frac{dP_0}{dQ^Y} = D(P||P^*) + D(P^*||Q), \end{aligned}$$

where we used the fact that all $P \in \boldsymbol{\mathcal{P}}$ have marginal $P^Y = P_0$. □

Remark 2. The law P^* is easily characterized in terms of the problem data P_0 and Q noticing that the marginal $P^{*Y} = P_0$ and the conditional $P^{*X|Y} = Q^{X|Y}$.

We now apply Lemma 1 to the case of normal laws and solve the first partial minimization. See also [2] for a different proof.

Proposition 3. *Let Q and P_0 be zero mean normal laws with strictly positive covariances $\Sigma \in \boldsymbol{\Sigma}$ and $\Sigma_0 \in \mathbb{R}^{n\times n}$ respectively. Then, $\min_{\Sigma' \in \boldsymbol{\Sigma}_0} D(\Sigma'||\Sigma)$ is attained by the zero mean normal law P^* with covariance*

$$\Sigma^* = \begin{pmatrix} \Sigma_0 & \Sigma_0\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{21}\Sigma_{11}^{-1}\Sigma_0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}(\Sigma_{11} - \Sigma_0)\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix} > 0.$$

Moreover,

$$D(\Sigma^*||\Sigma) = D(\Sigma_0||\Sigma_{11}).$$

Proof. This follows from Remark 2. A direct computation gives

$$\begin{aligned} \Sigma_{12}^* = \mathbb{E}_{P^*} XY^\top &= \mathbb{E}_{P^*}(\mathbb{E}_{P^*}[X|Y]Y^\top) \\ &= \mathbb{E}_{P^*}(\mathbb{E}_Q[X|Y]Y^\top) = \mathbb{E}_{P^*}(\Sigma_{21}\Sigma_{11}^{-1}YY^\top) \\ &= \Sigma_{21}\Sigma_{11}^{-1}\mathbb{E}_{P_0}YY^\top = \Sigma_{21}\Sigma_{11}^{-1}\Sigma_0. \end{aligned}$$

Likewise, we have

$$\begin{aligned} \Sigma_{22}^* = \mathbb{E}_{P^*} XX^\top &= \mathbb{C}\mathrm{ov}_{P^*}(X) \\ &= \mathbb{C}\mathrm{ov}_{P^*}(X|Y) + \mathbb{E}_{P^*}(\mathbb{E}_{P^*}[X|Y]\mathbb{E}_{P^*}[X|Y]^\top) \\ &= \mathbb{C}\mathrm{ov}_Q(X|Y) + \mathbb{E}_{P^*}(\mathbb{E}_Q[X|Y]\mathbb{E}_Q[X|Y]^\top) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \mathbb{E}_{P^*}(\Sigma_{21}\Sigma_{11}^{-1}Y(\Sigma_{21}\Sigma_{11}^{-1}Y)^\top) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \mathbb{E}_{P_0}(\Sigma_{21}\Sigma_{11}^{-1}YY^\top\Sigma_{11}^{-1}\Sigma_{12}) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{21}\Sigma_{11}^{-1}\Sigma_0\Sigma_{11}^{-1}\Sigma_{12}. \end{aligned}$$

Notice that, since $\Sigma > 0$ by assumption,

$$\Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^* = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} > 0$$

which, together with the assumption $\Sigma_0 > 0$, shows that $\Sigma^* > 0$.

The last relation, $D(\Sigma^*||\Sigma) = D(\Sigma_0||\Sigma_{11})$, reflects equation (7). □

3.2 The Second Partial Minimization Problem

In this section we turn to the second partial minimization problem. Here we minimize, for a given positive definite matrix $\Sigma \in \boldsymbol{\Sigma}$, the divergence $D(\Sigma||\Sigma_1)$ over $\Sigma_1 \in \boldsymbol{\Sigma}_1$.

Clearly this problem cannot have a unique solution in terms of the matrices H and Q. Indeed, if U is a unitary $k \times k$ matrix and $H' = HU$, $Q' = U^\top Q$, then $H'H'^\top = HH^\top$, $Q'^\top Q' = Q^\top Q$ and $H'Q' = HQ$. Nevertheless, the optimal matrices $HH^\top$, HQ and $Q^\top Q$ are unique, as we will see in Proposition 4. First we need to introduce some notation and conventions. If P is a positive definite matrix, we denote by $P^{1/2}$ any matrix satisfying $(P^{1/2})^\top(P^{1/2}) = P$, and by $P^{-1/2}$ its inverse. If M is any square matrix, we denote by $\Delta(M)$ the diagonal matrix

$$\Delta(M)_{ii} = M_{ii}.$$

Recall that we denote by $\Sigma(H, D, Q)$ a typical element of $\boldsymbol{\Sigma}_1$.

Proposition 4. *Given $\Sigma \in \boldsymbol{\Sigma}$ the $\min_{\Sigma_1 \in \boldsymbol{\Sigma}_1} D(\Sigma||\Sigma_1)$ is attained at a Σ_1^* such that $\Sigma_1 \in \boldsymbol{\Sigma}_1$ is solved by*

$$\begin{aligned} Q^* &= \Sigma_{22}^{1/2}, \\ H^* &= \Sigma_{12}\Sigma_{22}^{-1/2}, \\ D^* &= \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \end{aligned}$$

Thus the minimizing matrix $\Sigma_1^* = \Sigma(H^*, D^*, Q^*)$ becomes

$$\Sigma_1^* = \begin{pmatrix} \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{9}$$

Moreover, the Pythagorean law

$$D(\Sigma||\Sigma(H, D, Q)) = D(\Sigma||\Sigma_1^*) + D(\Sigma_1^*||\Sigma(H, D, Q)) \tag{10}$$

holds for any $\Sigma(H, D, Q) \in \boldsymbol{\Sigma}_1$, and therefore Σ_1^* is unique.

Proof. It is sufficient to show the validity of (10). We first compute

$$2D(\Sigma||\Sigma(H, D, Q)) - 2D(\Sigma||\Sigma_1^*).$$

It follows from Lemma A.2 that $|\Sigma(H, D, Q)| = |D| \times |Q^\top Q|$. In view of equation (3) the above difference becomes

$$\log|D| + \log|Q^\top Q| - \log|D^*| - \log|{Q^*}^\top Q^*| + \operatorname{tr}\big(\Sigma(H, D, Q)^{-1}\Sigma\big) - \operatorname{tr}\big(\Sigma_1^{*-1}\Sigma\big). \tag{11}$$

Using Corollary A.1, we compute

$$\Sigma(H, D, Q)^{-1} = \begin{pmatrix} D^{-1} & -D^{-1}HQ^{-\top} \\ -Q^{-1}H^\top D^{-1} & Q^{-1}(H^\top D^{-1}H + I)Q^{-\top} \end{pmatrix}, \tag{12}$$

and hence we get that

$$\begin{aligned} &\operatorname{tr}\big(\Sigma(H, D, Q)^{-1}\Sigma\big) = \operatorname{tr}\big(D^{-1}(\Sigma_{11} - HQ^{-\top}\Sigma_{21})\big) \\ &\quad + \operatorname{tr}\big(-Q^{-1}H^\top D^{-1}\Sigma_{12} + Q^{-1}(H^\top D^{-1}H + I)Q^{-\top}\Sigma_{22}\big) \\ &\quad = \operatorname{tr}\big(D^{-1}(\Sigma_{11} - 2HQ^{-\top}\Sigma_{21}) + Q^{-1}(H^\top D^{-1}H + I)Q^{-\top}\Sigma_{22}\big). \end{aligned} \tag{13}$$

Apply now Lemma A.2 to (9) and write $\Delta = \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$, to get

$$\Sigma_1^{*-1} = \Sigma(H^*, D^*, Q^*)^{-1} = \begin{pmatrix} \Delta^{-1} & -\Delta^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Delta^{-1} & \Sigma_{22}^{-1}\Sigma_{21}\Delta^{-1}\Sigma_{12}\Sigma_{22}^{-1} + \Sigma_{22}^{-1} \end{pmatrix}.$$

Therefore

$$\operatorname{tr}\big(\Sigma_1^{*-1}\Sigma\big) = \operatorname{tr}\big(\Delta^{-1} \times (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\big) + \operatorname{tr} I_k = \operatorname{tr}\big(\Delta^{-1}\Delta\big) + k = n + k. \tag{14}$$

Combining equations (11), (13), and (14), we find that

$$\begin{aligned} D(\Sigma||\Sigma(H,D,Q)) - D(\Sigma||\Sigma_1^*) = \\ \log|D| + \log|Q^\top Q| - \log|D^*| - \log|{Q^*}^\top Q^*| \\ + \mathrm{tr}\big(D^{-1}(\Sigma_{11} - HQ^{-\top}\Sigma_{21})\big) \\ + \mathrm{tr}\big(-Q^{-1}H^\top D^{-1}\Sigma_{12} + Q^{-1}(H^\top D^{-1}H + I)Q^{-\top}\Sigma_{22}\big) \\ - (n+k). \end{aligned} \tag{15}$$

We proceed with the computation of $2D(\Sigma_1^*||\Sigma(H,D,Q))$.

$$\begin{aligned} 2D(\Sigma(H^*,D^*,Q^*)||\Sigma(H,D,Q)) = \\ \log|D| + \log|Q^\top Q| - \log|D^*| - \log|{Q^*}^\top Q^*| - (n+k) \\ + \mathrm{tr}\big(\Sigma(H,D,Q)^{-1}\Sigma(H^*,D^*,Q^*)\big). \end{aligned} \tag{16}$$

Combining equations (9), (12), and $\mathrm{tr}\big(D^{-1}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \Delta)\big) = \mathrm{tr}\big(D^{-1}\Sigma_{11}\big)$, we obtain

$$\begin{aligned} \mathrm{tr}\big(\Sigma(H,D,Q)^{-1}\Sigma(H^*,D^*,Q^*)\big) = \mathrm{tr}\big(D^{-1}\Sigma_{11}\big) \\ - 2\mathrm{tr}\big(D^{-1}HQ^{-\top}\Sigma_{21}\big) + \mathrm{tr}\big(Q^{-1}(H^\top D^{-1}H + I)Q^{-\top}\Sigma_{22}\big). \end{aligned} \tag{17}$$

Insertion of (17) into (16) and a comparison with (15) yields the result. □

Remark 3. Notice that the matrix $H^*{H^*}^\top$ is strictly dominated by Σ_{11} (in the sense of positive matrices). This easily follows from $\Sigma_{11} - H^*{H^*}^\top = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} > 0$, and the assumption $\Sigma > 0$. By the same token $D^* > 0$.

3.3 The Link to the Original Problem

We now establish the connection between the lifted problem and the original Problem 1.

Proof of Proposition 2. Let $\Sigma_1 = \Sigma(H,D,Q)$ and denote by $\Sigma^* = \Sigma^*(\Sigma_1)$, the solution of the first partial minimization over $\boldsymbol{\Sigma}_0$. We have, for all $\Sigma' \in \boldsymbol{\Sigma}_0$,

$$\begin{aligned} D(\Sigma'||\Sigma_1) &\geq D(\Sigma^*||\Sigma_1) \\ &= D(\Sigma_0||HH^\top + D) \\ &\geq \min_{H,D} D(\Sigma_0||HH^\top + D), \end{aligned}$$

where we used Proposition 1 to write $\min$ on the RHS. It follows that

$$\inf_{\Sigma' \in \boldsymbol{\Sigma}_0, \Sigma_1 \in \boldsymbol{\Sigma}_1} D(\Sigma'||\Sigma_1) \geq \min_{H,D} D(\Sigma_0||HH^\top + D).$$

Conversely, let (H^*, D^*) be the minimizer of $(H,D) \mapsto D(\Sigma_0||HH^\top + D)$, pick an arbitrary invertible Q^*, and let $\Sigma^* = \Sigma(H^*, D^*, Q^*)$ be the corresponding element

in $\boldsymbol{\Sigma}_1$. Furthermore, let $\Sigma^{**} \in \boldsymbol{\Sigma}_0$ be the minimizer of $\Sigma \mapsto D(\Sigma||\Sigma^*)$ over $\boldsymbol{\Sigma}_0$. Then

$$\begin{aligned}\min_{H,D} D(\Sigma_0||HH^\top + D) &= D(\Sigma_0||H^*{H^*}^\top + D^*) \\ &\geq D(\Sigma^{**}||\Sigma^*) \\ &\geq \inf_{\Sigma' \in \boldsymbol{\Sigma}_0, \Sigma_1 \in \boldsymbol{\Sigma}_1} D(\Sigma'||\Sigma_1),\end{aligned}$$

which shows the opposite inequality. Finally, to show that we can replace the infima with minima also in the lifted problem, notice that (see Proposition 3) $D(\Sigma^{**}||\Sigma^*) = D(\Sigma_0||H^*{H^*}^\top + D^*)$. □

4 Alternating Minimization Algorithm

In this section we combine the two partial minimization problems above to derive an iterative algorithm for Problem 1. It turns out that this algorithm is also instrumental in proving the existence of a solution to Problem 1.

4.1 The Algorithm

We suppose that the given matrix Σ_0 is strictly positive definite. Pick the initial values H_0, D_0, Q_0 such that H_0 is of full rank, $D_0 > 0$ is diagonal, Q_0 and $H_0H_0^\top + D_0$ are invertible.

At the t-th iteration the matrices H_t, D_t and Q_t are available. Start solving the first partial minimization problem with $\Sigma = \Sigma(H_t, D_t, Q_t)$. Use the resulting matrix as data for the second partial minimization, the solution of which gives the update rules

$$\begin{aligned}Q_{t+1} = \Big(&Q_t^\top Q_t - Q_t^\top H_t^\top (H_tH_t^\top + D_t)^{-1} H_t Q_t \\ &+ Q_t^\top H_t^\top (H_tH_t^\top + D_t)^{-1} \Sigma_0 (H_tH_t^\top + D_t)^{-1} H_t Q_t\Big)^{1/2}, \qquad (18)\end{aligned}$$

$$H_{t+1} = \Sigma_0 (H_tH_t^\top + D_t)^{-1} H_t Q_t Q_{t+1}^{-1}, \qquad (19)$$

$$D_{t+1} = \Delta(\Sigma_0 - H_{t+1}H_{t+1}^\top). \qquad (20)$$

In (18) there is some freedom in computing the square root that determines Q_{t+1}. Properly choosing the square root will result in the disappearance of Q_t from the algorithm. This is an attractive feature, since Q_t only serves as an auxiliary variable. One can write the RHS of equation (18), before taking the square root, as

$$Q_t^\top (I - H_t^\top (H_tH_t^\top + D_t)^{-1}(H_tH_t^\top + D_t - \Sigma_0)(H_tH_t^\top + D_t)^{-1} H_t) Q_t$$

and denoting

$$R_t = I - H_t^\top (H_tH_t^\top + D_t)^{-1}(H_tH_t^\top + D_t - \Sigma_0)(H_tH_t^\top + D_t)^{-1} H_t \qquad (21)$$

a possible square root is given by

$$R_t^{1/2} Q_t.$$

Notice that R_t only involves the iterates H_t and D_t. The update equation (18) can therefore be rewritten as

$$H_{t+1} = \Sigma_0 (H_t H_t^\top + D_t)^{-1} H_t R_t^{-1/2}. \tag{22}$$

The final version of the algorithm is given by equations (20),(21), and (22) which, for clarity, we present as

Algorithm 1

$$R_t = I - H_t^\top (H_t H_t^\top + D_t)^{-1} (H_t H_t^\top + D_t - \Sigma_0)(H_t H_t^\top + D_t)^{-1} H_t, \tag{23}$$

$$H_{t+1} = \Sigma_0 (H_t H_t^\top + D_t)^{-1} H_t R_t^{-1/2}, \tag{24}$$

$$D_{t+1} = \Delta(\Sigma_0 - H_{t+1} H_{t+1}^\top). \tag{25}$$

In order to avoid taking a square root at each step one can introduce the matrices $K_t = H_t Q_t$ and $P_t = Q_t^T Q_t$ and write the updates for K_t and P_t. Equations (18), (19), and (20) easily give

Algorithm 2

$$\begin{aligned} K_{t+1} &= \Sigma_0 (K_t P_t^{-1} K_t^\top + D_t)^{-1} K_t, \\ P_{t+1} &= P_t - K_t^\top (K_t P_t^{-1} K_t^\top + D_t)^{-1} (K_t P_t^{-1} K_t^\top + D_t - \Sigma_0)(K_t P_t^{-1} K_t^\top + D_t)^{-1} K_t, \\ D_{t+1} &= \Delta(\Sigma_0 - K_{t+1} P_{t+1}^{-1} K_{t+1}^\top). \end{aligned} \tag{26}$$

After the final iteration, the T-th say, one can take $H_T = K_T Q_T^{-1}$, where Q_T is a square root of P_T.

Notice that in both Algorithm 1 and 2 it is required to invert $n \times n$ matrices (like e.g. $(H_t H_t^\top + D_t)^{-1}$). Applying corollary A.1 one gets $(H_t H_t^\top + D_t)^{-1} H_t = D_t^{-1} H_t (I + H_t^\top D_t^{-1} H_t)$. Hence, we can replace e.g. (22) with

$$H_{t+1} = \Sigma_0 D_t^{-1} H_t (I + H_t^\top D_t^{-1} H_t)^{-1} R_t^{-1/2}. \tag{27}$$

By the same token one can write

$$K_{t+1} = \Sigma_0 D_t^{-1} K_t (P_t + K_t^\top D_t^{-1} K_t)^{-1} P_t$$

to replace (26).

Some properties of the algorithm are summarized in the next proposition.

Proposition 5. *For Algorithm 1 the following hold for all t.*

(a) $D_t > 0$ and $(D_t)_{ii} \leq (\Sigma_0)_{ii}$.
(b) R_t is invertible.
(c) If H_0 is of full column rank, so is H_t.
(d) $H_t H_t^\top \leq \Sigma_0$.

(e) If $\Sigma_0 = H_0 H_0^\top + D_0$ then the algorithm stops.

(f) The objective function decreases at each iteration. More precisely, let $\Sigma_{0,t}$ be the solution of the first partial minimization with data $\Sigma_t = \Sigma(H_t, D_t, Q_t)$. Then

$$D(\Sigma_0||H_{t+1}H_{t+1}^\top) - D(\Sigma_0||H_t H_t^\top) = -\Big(D(\Sigma_{t+1}||\Sigma_t) + D(\Sigma_{0,t}||\Sigma_{0,t+1})\Big).$$

(g) The limit points (H, D) of the algorithm satisfy the relations

$$H = (\Sigma_0 - HH^\top)D^{-1}H,$$
$$D = \Delta(\Sigma_0 - HH^\top).$$

Proof. (a) This follows from Remark 3.

(b) Use the identity $I - H_t^\top(H_t H_t^\top + D_t)^{-1}H_t = (I + H_t^\top D_t^{-1} H_t)^{-1}$ and the assumption $\Sigma_0 > 0$.

(c) Use the assumption $\Sigma_0 > 0$, (a), and (b).

(d) Again from Remark 3 and the construction of the algorithm as a combination of the two partial minimization problems.

(e) This is a triviality upon noticing that one can take $R_t = I$ in this case.

(f) It follows from a concatenation of Lemma 1 and Proposition 4. Notice that we can express the decrease as the sum of two I-divergences, since the Pythagorean law holds for both partial minimizations.

(g) We consider Algorithm 2 first. Assume that all variables converge. Then, from (26), the limit points K, P, D satisfy the relation $K = \Sigma_0 D^{-1}K(P + K^\top D^{-1}K)^{-1}P$. Postmultiplication by $P^{-1}(P + K^\top D^{-1}K)$ yields, after rearranging terms, $K = (\Sigma_0 - KP^{-1}K^\top)D^{-1}K$. Let now Q be a square root of P and $H = KQ^{-1}$ to get the first relation. The rest is trivial. □

4.2 Proof of Proposition 1

Let D_0 and H_0 be arbitrary and perform one step of the algorithm to get matrices D_1 and H_1. It follows from Proposition 5 that $D(\Sigma_0||H_1H_1^\top + D_1) \leq D(\Sigma_0||H_0H_0^\top + D_0)$. Moreover, $H_1H_1^\top \leq \Sigma_0$ and $D_1 \leq \Delta(\Sigma_0)$. Hence the search for a minimum can be confined to the set of matrices (H, D) satisfying $HH^\top \leq \Sigma_0$ and $D \leq \Delta(\Sigma_0)$. Next, we claim that it is also sufficient to restrict the search for a minimum to all matrices (H, D) such that $HH^\top + D \geq \varepsilon I$ for some sufficiently small $\varepsilon > 0$. Indeed, if the last inequality is violated, then $HH^\top + D$ has an eigenvalue less than ε. Write the Jordan decompositions $HH^\top + D = U\Lambda U^\top$, and let $\Sigma_U = U^\top \Sigma_0 U$. Then $D(\Sigma_0||HH^\top + D) = D(\Sigma_U||\Lambda)$, as one easily verifies. Denoting by λ_i the eigenvalues of $HH^\top + D$ and letting σ_{ii} be the diagonal elements of Σ_U, we can write $D(\Sigma_U|\Lambda) = -\frac{1}{2}\log|\Sigma_U| + \frac{1}{2}\sum_i \log\lambda_i - \frac{n}{2} + \frac{1}{2}\sum_i \frac{\sigma_{ii}}{\lambda_i}$. Let λ_{i_0} be a minimum eigenvalue and take ε smaller than the minimum of all σ_{ii}, which is positive, since Σ_0 is strictly positive definite. Then the contribution for $i = i_0$ in the summation to the divergence $D(\Sigma_U||\Lambda)$ is at least $\log\varepsilon + 1$, which tends to infinity for $\varepsilon \to 0$. This proves the claim. So, we have shown that a minimizing pair (H, D) has to satisfy $HH^\top \leq \Sigma_0$, $D \leq \Delta(\Sigma_0)$, and $HH^\top + D \geq \varepsilon I$, for some $\varepsilon > 0$. In other words we have to minimize the I-divergence over a compact set on which it is clearly continuous. This proves Proposition 1. □

References

1. T.W. Anderson (1984), *An introduction to multivariate statistical analysis*, Second ed., Wiley.
2. E. Cramer (1998), Conditional iterative proportional fitting for Gaussian distributions, *J. Multivariate Analysis*, **65(2)**, 261–276.
3. E. Cramer (2000), Probability measures with given marginals and conditionals: I-projections and conditional iterative proportional fitting, *Statistics & Decisions*, **18(3)**, 311–329.
4. I. Csiszár and G. Tusnády (1984), Information geometry and alternating minimization procedures, *Statistics & Decisons, supplement issue* **1**, 205-237.
5. L. Finesso and G. Picci (1984), Linear statistical models and stochastic realization theory. *Analysis and optimization of systems, Part 1 (Nice, 1984)*, 445–470, Lecture Notes in Control and Inform. Sci., 62, Springer, Berlin.
6. L. Finesso and P. Spreij (2006), Nonnegative matrix factorization and I-divergence alternating minimization, *Linear Algebra and its Applications*, **416**, 270–287.

A Appendix

For ease of reference we collect here some standard formulas for the normal distribution and some matrix algebra.

A.1 Multivariate Normal Distribution

Let $(X^\top, Y^\top)^\top$ be a zero mean normal vector with covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}.$$

Assume that Σ_{YY} is invertible. The conditional law of X given Y is normal with $\mathbb{E}[X|Y] = \Sigma_{XY}\Sigma_{YY}^{-1}Y$ and

$$\mathbb{C}\mathrm{ov}[X|Y] = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}. \tag{28}$$

A.2 Partitioned Matrices

Lemma 2. *Let A, D be square matrices. Assume invertibility where required.*

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix} = \begin{pmatrix} I & CD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - CD^{-1}B & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}B & I \end{pmatrix},$$

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - BA^{-1}C \end{pmatrix} \begin{pmatrix} I & A^{-1}C \\ 0 & I \end{pmatrix},$$

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - CD^{-1}B)^{-1} & -(A - CD^{-1}B)^{-1}CD^{-1} \\ -D^{-1}B(A - CD^{-1}B)^{-1} & D^{-1}B(A - CD^{-1}B)^{-1}CD^{-1} + D^{-1} \end{pmatrix}.$$

Corollary 1

$$(D - BAC)^{-1} = D^{-1} + D^{-1}B(A^{-1} - CD^{-1}B)^{-1}CD^{-1}.$$

Proof. For Lemma 2 a check will suffice. The Corollary follows using the two decompositions of the Lemma with A replaced by A^{-1} and comparing the two expressions of the lower right block of the inverse matrix. □