Sparse factor analysis
via likelihood and ℓ1-regularization

Lipeng Ning and Tryphon T. Georgiou

Abstract—In this note we consider the basic problem to identify linear relations in noise. We follow the viewpoint of factor analysis (FA) where the data is to be explained by a small number of independent factors and independent noise. Thereby an approximation of the sample covariance is sought which can be factored accordingly. An algorithm is proposed which weighs in an ℓ1-regularization term that induces sparsity of the linear model (factor) against a likelihood term that quantifies distance of the model to the sample covariance. The algorithm compares favorably against standard techniques of factor analysis. Their performance is compared first by simulation, where ground truth is available, and then on stock-market data where the proposed algorithm gives reasonable and sparser models.

I. INTRODUCTION

We are interested in a most basic problem where linear models are sought from noisy data [10]. The salient feature of the problem, however, is that it involves typically a large number of variables. This of great importance in many areas in science, engineering, but most notably in econometrics and economic forecasting [4].

Consider the (column) vector

\[ x = [x_1, \ldots, x_n]' \in \mathbb{R}^n \]

of real random variables \( x_1, \ldots, x_n \), and a set of independent realizations of \( x(1), \ldots, x(m) \) of the vector \( x \) making up

\[ X = [x(1), \ldots, x(m)] \in \mathbb{R}^{n \times m}. \]

We assumed that these have zero mean and we set

\[ \Sigma_{\text{sample}} = \frac{1}{m} XX' \]

to be the sample covariance.

From this point, various modeling postulates can be considered. Each makes different assumption about the nature of noise. Typically, one assumes that

\[ x = \hat{x} + \tilde{x} \]

where \( \hat{x} \) represents a noise-free variable and \( \tilde{x} \) represents noise. Then, \( \hat{x} \) and \( \tilde{x} \) are assumed uncorrelated zero-mean (vectorial) random variables. The classical Frisch model postulates that

\[ \hat{\Sigma} := E(\hat{x}\hat{x}') \]

is diagonal whereas principle component analysis (PCA) assumes that \( \Sigma \) has no structure whatsoever, but that the variance \( \text{trace}(\Sigma) \) is small.

Regarding the structure of the noise-free variables, it is assumed that these satisfy between them a number of linear relations. Thus, the rank of the noise-free covariance

\[ \hat{\Sigma} = E(\hat{x}\hat{x}') \]

must be less than \( n \). This \( \text{rank}(\hat{\Sigma}) \) is the number of “principle components” or “factors” which are needed to explain the data.

In the PCA literature, the rank is conveniently chosen according to a “break point” of the singular values of \( \Sigma_{\text{sample}} \), while in the context of the Frisch problem, it is natural to seek \( \Sigma \) of minimal rank [10]–a problem which to large degree is still open. Factor analysis on the other hand focuses not not only on the number of independent components, which need to be small, but also on the structure of the postulated noise-free covariance \( \Sigma \). More specifically, it is assumed that

\[ X = FV + \hat{X} \]

and that the noise-free data \( \hat{X} = FV \) is modeled by a constant coefficient matrix \( F \), which is required to be sparse, and a matrix

\[ V = [v(1), \ldots, v(m)], \]

where \( v(t) = [v_1(t), \ldots, v_r(t)]' \), \( t = 1, \ldots, m \) represent independent realizations of a set of random variables. The variables \( v_j \) are called the factors and \( F \) is the factor loading in factor analysis.

The underlying rationale for sparse factor loadings is so that each random variable \( \hat{x}_j \) is accounted for by superimposing the effect of a small number of factors \( v_j \). The assumption on \( \hat{x} \) is again that the entries are independent and hence that the covariance \( \hat{\Sigma} \) is diagonal. Further, without loss of generality, the covariance of \( v \) can be taken to be the identity. Thus the basic problem in factor analysis is as follows.

Problem: Given an \( n \times n \) sample covariance \( \Sigma_{\text{sample}} \) and an integer \( r < n \), determine a diagonal covariance \( \Sigma \) and a sparse coefficient matrix \( F \in \mathbb{R}^{n \times r} \) such that

\[ \Sigma_{\text{sample}} \simeq FF' + \hat{\Sigma}. \] (1)

The purpose of this paper is to propose an algorithm for obtaining such a decomposition. The solution is not optimal, i.e., it is neither sparsest nor of least rank. Attaining either of these specs is a formidable objective. Instead, we consider a likelihood function together with a regularization term that promotes sparsity. The algorithm aims to approximate extrema of such a functional.
II. ALGORITHM

Suppose that the data is drawn from a Gaussian distribution with covariance $\Sigma$. In the absence of any constraints the negative-log-likelihood

$$\frac{m}{2} \left( \log \det(\Sigma) + \text{trace}(\Sigma_{\text{sample}} \Sigma^{-1}) + n \log(2\pi) \right)$$

is minimal at $\Sigma = \Sigma_{\text{sample}}$, while in general, it can be taken to quantify distance between $\Sigma$ and $\Sigma_{\text{sample}}$.

Given the sample covariance $\Sigma_{\text{sample}}$, our interest is in obtaining a nearby value for $\Sigma$ which admits a factorization as in (1) with a sparse factor $F$. The sparsity of $F$, often denoted by $|F|_0$, is the number of non-zero entries. It is known that the sum of the absolute values of its entries, namely $|F|_1$, is a good surrogate for sparsity. The use of the $\ell_1$ norm for penalty to promote sparsity has a history in statistics, but was brought to prominence only recently, after a series of deep studies by Candès, Tao, Romberg, Donoho, Elad, and others.

Thus, for our purposes, we propose as a compromise between fit and sparsity, a linear combination of the likelihood function and a weighted $\ell_1$-norm of the corresponding factor. This linear combination of likelihood and $\ell_1$-cost is thought of as a function of the entries $F$, $\tilde{\Sigma}$ in the decomposition

$$\Sigma = FF' + \tilde{\Sigma}$$

with $\tilde{\Sigma} \geq 0$ and diagonal. Thus, our problem can be expressed as follows:

$$\min_{F, \tilde{\Sigma}} \left\{ \log \det(FF' + \tilde{\Sigma}) + \text{trace}(\Sigma_{\text{sample}}(FF' + \tilde{\Sigma})^{-1}) \right\}$$

$$\quad + \lambda |F|_1 | \tilde{\Sigma} \geq 0, \tilde{\Sigma} \text{ diagonal} \}. \quad (2)$$

The weight $\lambda$ affects sparsity of the factor $F$.

Problem (2) is not convex. Since

$$(FF' + \tilde{\Sigma})^{-1} = E - GG'$$

with

$$E = \tilde{\Sigma}^{-1}, \quad (3a)$$

$$G = \tilde{\Sigma}^{-1} FM^{-1}, \quad (3b)$$

$$M = (I + F' \tilde{\Sigma}^{-1} F)^{-\frac{1}{2}}, \quad (3c)$$

then (2) becomes

$$\min_{G, E} \left\{ - \log \det(E - GG') + \text{trace}(\Sigma_{\text{sample}}(E - GG')) \right\}$$

$$\quad + \lambda \|E^{-1}GM\|_1 | E - GG' \geq 0, E \text{ is diagonal} \}. \quad (4)$$

For $G = G_k + \delta G$, the first order approximation of $GG'$ is

$$\delta G G_k' + G_k \delta G' + G_k G_k' =: [GG']_{k, \delta G}. \quad (5)$$

Likewise, let $E = E_k + \delta E$ where $G_k, E_k$ for $k = 1, 2, \ldots$ represents iteratively obtained values for $G$ and $E$ (as in (6b-6c) below), respectively. Also $M$ and $F$ will be approximated iteratively. Since, from (3b) $F = E^{-1}GM$ we approximate $F$ with the first order perturbation

$$E_k^{-1}(G_k + \delta G)M_k,$$

where $M_k$ represents the value of $M$ from (3c) at $E_k$ and $G_k$.

In this expression, we neglect the linear perturbation terms corresponding to $E$ and $M$. Our reasoning is as follows: first, $E$ is diagonal and neglecting corrections will not affect the sparsity pattern of $F$ while on the other hand the linear perturbation term involves $E_k^{-2}$ which causes the iteration to be numerically sensitive. Then again, the linear perturbation in $M$ is rather complicated (given in Appendix B) and the neglected term is not affecting the convergence claimed in the proposition below. Define

$$g_{G_k, E_k}(\delta G, \delta E) = - \log \det(E_k + \delta E - [GG']_{k, \delta G})$$

$$\quad + \text{trace}(\Sigma_{\text{sample}}(E + \delta E - [GG']_{k, \delta G}))$$

$$\quad + \lambda \|E_k^{-1}(G_k + \delta G)M_k\|_1$$

which is convex in $\delta G$ and $\delta E$ and approximates (4) at $G_k$ and $E_k$. Moreover, the constraint $E - GG' \geq 0$ is equivalent to

$$\begin{bmatrix} E & G \\ G' & I \end{bmatrix} \geq 0.$$ 

Hence, we seek minima of (4) by solving

$$(\hat{\delta} G, \hat{\delta} E) = \arg \min \left\{ g_{G_k, E_k}(\delta G, \delta E) : \right\}$$

$$\quad \begin{bmatrix} E_k + \delta E & G_k + \delta G \\ G_k' + \delta G' & I \end{bmatrix} \geq 0,$$

$$\text{and } \delta E \text{ diagonal} \}. \quad (6a)$$

We choose step size $\alpha \in [0, 1]$ such that with

$$G_{k+1} = G_k + \alpha \delta G \quad (6b)$$

$$E_{k+1} = E_k + \alpha \delta E \quad (6c)$$

the following inequality holds

$$g_{G_{k+1}, E_{k+1}} < g_{G_k, E_k} - \sigma. \quad (6d)$$

Here, the constant $\sigma > 0$ determines a stopping criterion. Also, we have simplified the notation by setting

$$g_{G_k, E_k} := g_{G_k, E_k}(0, 0).$$

If (6d) holds, the step size $\alpha$ is accepted, otherwise let $\alpha = \frac{1}{2} \alpha$. If the step size is smaller than a preselected $\epsilon > 0$, the iterations can be terminated.

**Proposition:** If $\lambda = 0$, then at point $G_k, E_k$, $(\hat{\delta} G, \hat{\delta} E)$ is descent direction for (4).

**Proof:** Since $\delta G = \delta E = 0$ satisfy the LMI in (6),

$$g_{G_k, E_k}(\delta G, \delta E) \leq g_{G_k, E_k}.$$ 

Moreover, $g_{G_k, E_k}(\delta G, \delta E)$ is convex in $\delta G$ and $\delta E$, so

$$g_{G_k, E_k}(\epsilon \delta G, \epsilon \delta E) \leq g_{G_k, E_k}$$

for all $\epsilon \in [0, 1]$. If $\epsilon$ is small, the difference between the above two is of order $\epsilon$ or higher. When $\lambda = 0$, because of
(5), the difference between \( g_{G_k,E_k}(\hat{\delta}_G, \hat{\delta}_E) \) and the objective function in (4) is of order \( \varepsilon^2 \) or higher. So \( (\hat{\delta}_G, \hat{\delta}_E) \) is descent direction for (4) at \( E_k, G_k \).

If \( \lambda > 0 \) and is small enough, the decreased amount of the first two terms dominates the change of the last term. So the pair \( (\hat{\delta}_G, \hat{\delta}_E) \) is still a descent direction. Since the objective function (4) is bounded below, then the algorithm will at least converge to a local minimum.

The algorithm can be initialized using suitable starting values for \( \hat{\Sigma}_0 \) and \( F_0 \), and then setting

\[
M_0 = (I + F_0 \hat{\Sigma}_0^{-1} F_0)^{-\frac{1}{2}}
\]
\[
E_0 = \hat{\Sigma}_0^{-1}
\]
\[
G_0 = \hat{\Sigma}_0^{-1} F_0 M_0^{-1}.
\]

A starting value for \( F_0 \) can be chosen to contain the eigenvectors of \( \hat{\Sigma}_\text{sample} \) corresponding to the largest \( r \) eigenvalues, scaled by the square root of corresponding eigenvalues, and \( \hat{\Sigma}_0 \) may be set equal to the diagonal \(^1\) of \( \hat{\Sigma}_0 - F_0 F_0' \).

III. BACKGROUND AND THE VARIMAX CRITERION

The problem to identify linear relations in data has its roots at least as far back as in the work of Gauss. “Least squares” has been a workhorse in engineering ever since. Early in the 20th century, following Ragnar Frisch, statisticians laid down alternative assumptions on the noise model and sought to understand the impact of such concepts on modeling.

There are several schools of thought. Most prominently, principal component analysis (PCA) which is based on the fact that singular value decomposition (SVD) allows for an exact and computationally simple analysis of data and covariance according to the hypothesis that noise has no structure while the signal-to-noise ratio is significant. In parallel, in disciplines such as psychometrics and econometrics where data is often dominated by noise [11], [16], assuming a more detailed noise-model is essential. Reasonable hypotheses often allow for more accurate models, albeit at a cost of an often computationally intractable problem. The main schools are that of Factor Analysis (FA) and Errors in variables (EIV). FA is based on the assumption of the independence of noise whereas EIV allow for more sophisticated models and is a broader research area [13], [14]. The assumption of independent noise components is natural in several applications in signal analysis and system identification, but most importantly in financial data. In [2] the sparse factor model was assumed in the study of gene expression genomics where a Bayesian work was introduced. In [3] the same objective function, penalized maximum likelihood function, was considered for the sparse factor analysis problem and a generalized expectation maximization algorithm was proposed. A recent study in [17] to achieve sparse PCA blares the distinction as it assumes a noise model with a diagonal structure. Further, in [17] a similar weighing of a likelihood function together with an \( \ell_1 \)-penalty is being proposed. See also [18] for an alternative view to sparse PCA.

Dynamic modeling can also be formulated in a variety of ways [1], [4]. One way that this can be achieved is by seeking linear relations between time-shifted copies of time-series data and will not be discussed further. The problem to sparsify factor loadings has been a central issue in FA. A recent approach is to search for an \( r \times r \) rotation matrix \( R \), in conjunction with a search for the \( n \times r \) factor loading \( F \), so that the product \( FR \) is sparse. Cumberose as it may seem, this viewpoint forms the basis of a technique presented in [9], [12]. The rotation matrix \( R \) is sought to maximize the following Varimax criterion:

\[
\sum_{j=1}^{r} \left( \sum_{i=1}^{n} (FR)_{ij}^4 - \frac{1}{n} \sum_{i=1}^{n} (FR_{ij})^2 \right).
\]

This expression is precisely the sum of variances of the squares of the elements of \( FR \). The rationale for seeking extrema of this expression rests on the observation that the Frobenius norm of \( FR \) is not affected by \( R \), which is a rotation matrix. Then, if this is chosen so as to render the variance of the squares of the entries large, \( FR \) will necessarily have most entries small. This criterion is now standard and is included as part of Matlab in the subroutine \texttt{factoran} which is widely used for factor analysis.

IV. EXAMPLES

Example 1: First we compare the performance of the proposed method against \texttt{factoran} (which uses the Varimax criterion) using simulation. To this end, we generate a random factor loading matrix \( F \in \mathbb{R}^{100 \times 10} \) with only two non-zero entries per row, and a realization \( V \in \mathbb{R}^{10 \times 400} \) of random factors from a multivariate normal distribution with covariance equal to the identity. Figure 1 displays a color-coded representation of the entries of \( F \), side by side with the entries of the estimated factor loadings using our algorithm (second column) and one produced by \texttt{factoran} (third column). To highlight the difference we display in Figure 2 the same, as a binary plot, thresholding at a low amplitude. This shows an almost perfect agreement of the result of our algorithm with the “ground truth.” Finally, Figure 3 shows the relative values of entries across one of the columns of the factor loading matrix \( F \) (5th column) and compares, likewise, the “ground truth” with the result of the algorithm presented in this paper and the output of \texttt{factoran}. It is seen that both algorithms are consistent.

Example 2: We process time-series data corresponding to 30 different stocks taken from 4 different sectors. The stocks and sectors are listed in the Appendix A. The data are taken over a period of four weeks sampled at 2 minute intervals.
The same database has been used in [15]. We compute the correlation matrix which is shown in Figure 4 in color-coded format. From the figure it is possible to discern strong correlation between certain stocks. This in turn may suggest a possible common underlying factor. Otherwise noise may be the dominant component. We compute factor loadings using the method that was proposed earlier and compare with the factor loadings obtained using standard factor analysis (factoran routine in Matlab). These are displayed in Figure 5, again in a color-coded format. It is evident that the method that we propose gives sparser factor loadings in this example as well.

In general, we observe that both our algorithm and the factoran are largely consistent. In this example there is no “ground truth.” Inspection of the sample covariance and further analysis can suggest whether values in the factor loadings seem appropriate, although such a claim is difficult to substantiate. Yet, it is apparent from Figure 5 that our algorithm relies less on small values in the factor loading to explain the data. Hence, the factor loading matrix $F$ is sparser.

ACKNOWLEDGMENT

The financial-stock data used in Example 2 was provided to us by Dr. Donatello Materassi.

REFERENCES


Given $A$ and $\Delta$,

$$e^{A+\Delta} = e^A + \int_0^1 e^{(1-\tau)A} \Delta e^{\tau A} d\tau + o(\|\Delta\|),$$

(7a)

$$\log(A + \Delta) = \log A + \int_0^\infty (A + \tau I)^{-1} \Delta (A + \tau I)^{-1} d\tau$$

$$+ o(\|\Delta\|),$$

(7b)

see e.g. [7]. Denote $\Delta := G_k^t E_k G_k - G' E G$ and write

$$M = (I - G'E) - \frac{1}{2} = (I - G_k^t E_k G_k + \Delta) - \frac{1}{2}$$

$$= e^{-\frac{1}{2} \log(I - G_k^t E_k G_k + \Delta)}.$$

Then from (7)

$$M = M_k - \frac{1}{2} \int_0^1 \int_0^\infty M_k^{1-\lambda} (I - G_k^t E_k G_k + \tau I)^{-1} \Delta \times$$

$$(I - G_k^t E_k G_k + \tau I)^{-1} M_k^{1}\Delta d\tau d\lambda + o(\|\Delta\|).$$

Appendix A

<table>
<thead>
<tr>
<th>Name</th>
<th>Sector</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Utilities</td>
<td>Electric Utilities</td>
</tr>
<tr>
<td>DTE</td>
<td>Utilities</td>
<td>Electric Utilities</td>
</tr>
<tr>
<td>DUK</td>
<td>Utilities</td>
<td>Electric Utilities</td>
</tr>
<tr>
<td>ED</td>
<td>Utilities</td>
<td>Electric Utilities</td>
</tr>
<tr>
<td>EIX</td>
<td>Utilities</td>
<td>Electric Utilities</td>
</tr>
<tr>
<td>ADI</td>
<td>Technology</td>
<td>Semiconductors</td>
</tr>
<tr>
<td>ALTR</td>
<td>Technology</td>
<td>Semiconductors</td>
</tr>
<tr>
<td>AMAT</td>
<td>Technology</td>
<td>Semiconductors</td>
</tr>
<tr>
<td>AMD</td>
<td>Technology</td>
<td>Semiconductors</td>
</tr>
<tr>
<td>BRCM</td>
<td>Technology</td>
<td>Semiconductors</td>
</tr>
<tr>
<td>CMA</td>
<td>Financial</td>
<td>Regional Banks</td>
</tr>
<tr>
<td>FHN</td>
<td>Financial</td>
<td>Regional Banks</td>
</tr>
<tr>
<td>FITB</td>
<td>Financial</td>
<td>Regional Banks</td>
</tr>
<tr>
<td>HBAN</td>
<td>Financial</td>
<td>Regional Banks</td>
</tr>
<tr>
<td>KEY</td>
<td>Financial</td>
<td>Regional Banks</td>
</tr>
<tr>
<td>MER</td>
<td>Financial</td>
<td>Investment Services</td>
</tr>
<tr>
<td>MS</td>
<td>Financial</td>
<td>Investment Services</td>
</tr>
<tr>
<td>NYX</td>
<td>Financial</td>
<td>Investment Services</td>
</tr>
<tr>
<td>SCHW</td>
<td>Financial</td>
<td>Investment Services</td>
</tr>
<tr>
<td>TROW</td>
<td>Financial</td>
<td>Investment Services</td>
</tr>
<tr>
<td>APA</td>
<td>Energy</td>
<td>Oil &amp; Gas Operations</td>
</tr>
<tr>
<td>APC</td>
<td>Energy</td>
<td>Oil &amp; Gas Operations</td>
</tr>
<tr>
<td>CHK</td>
<td>Energy</td>
<td>Oil &amp; Gas Operations</td>
</tr>
<tr>
<td>DYN</td>
<td>Energy</td>
<td>Oil &amp; Gas Operations</td>
</tr>
<tr>
<td>ABK</td>
<td>Financial</td>
<td>Insurance (Prop.&amp; Casualty)</td>
</tr>
<tr>
<td>ACE</td>
<td>Financial</td>
<td>Insurance (Prop.&amp; Casualty)</td>
</tr>
<tr>
<td>AIG</td>
<td>Financial</td>
<td>Insurance (Prop.&amp; Casualty)</td>
</tr>
<tr>
<td>ALL</td>
<td>Financial</td>
<td>Insurance (Prop.&amp; Casualty)</td>
</tr>
<tr>
<td>CB</td>
<td>Financial</td>
<td>Insurance (Prop.&amp; Casualty)</td>
</tr>
</tbody>
</table>

Appendix B

Then from (7)

$$M = M_k - \frac{1}{2} \int_0^1 \int_0^\infty M_k^{1-\lambda} (I - G_k^t E_k G_k + \tau I)^{-1} \Delta \times$$

$$(I - G_k^t E_k G_k + \tau I)^{-1} M_k^{1}\Delta d\tau d\lambda + o(\|\Delta\|).$$