# Sparsest Factor Analysis of Gene Expression Data

Kohei Adachi* and Nickolay T. Trendafilov†

*Abstract* — **We consider factor analysis (FA) of an *n*-observations × *p*-genes data matrix with *p* >> *n*. For such high-dimensional data, FA produces factor loading matrix with too many rows to be easily interpreted. To deal with this difficulty, we consider an FA procedure subject to the loading matrix being sparsest in that each row of the matrix has only one nonzero loading. As such a sparsest FA procedure, a least squares matrix-decomposition approach has been presented by the authors, but it is not feasible for high-dimensional case. In this paper, a maximum-likelihood sparsest FA procedure is proposed which is feasible for the data with *p* >> *n*. Its key point is using the EM algorithm in which *n* > *p* is indispensable. The proposed method is applied for a gene expression data set.**

**Keyword:** *High-dimensional data; The sparsest loadings; EM algorithms*

## 1 Introduction

In factor analysis (FA), the variations of $p$ observed variables are assumed to be explained by the two types of unobserved factors called common factors and unique factors, with their uncorrelated mutually. Here, the number of the common factors, which we denote as $m$, are supposed to be less than that of observed variables, i.e., $m < p$, and those factors serve for explaining the variations of all $p$ variables. On the other hand, the number of the unique factors equals that of observed variables ($p$) and the factors have the one-to-one relationship to $p$ variables: each of unique factors explains specifically the variation of the corresponding variable.

Using $\mathbf{x}$ ($p \times 1$) for a $p$-variate vector with its expectation E$[\mathbf{x}]$ equaling the zero vector $\mathbf{0}_p$, the FA model is expressed as $\mathbf{x} = \mathbf{\Lambda f} + \mathbf{e}$. Here, $\mathbf{f}$ is an $m \times 1$ common factor vector with $m < p$, $\mathbf{e}$ is a $p \times 1$ unique factor vector, and $\mathbf{\Lambda}$ is a $p \times m$ factor loading matrix describing the relationships between the $p$ variables and $m$ factors. The $p \times p$ covariance matrix $\mathbf{\Psi}$ for $\mathbf{e}$ is assumed to be a diagonal matrix, whose diagonal elements are called unique variances, as they stand for the amount of variances of variables accounted for by unique factors. In FA, $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and factor correlation matrix $\mathbf{\Phi}$ are estimated from the $n$-observations × $p$-variables data matrix $\mathbf{X}$ whose rows are the realizations of $\mathbf{x}'$.

In this paper, we focus on gene-expression data matrices $\mathbf{X}$ with $p >> n$, i.e., much more variables (i.e., genes) than observations. For such a case, the resulting $p \times m$ loading matrix $\mathbf{\Lambda}$ has too many rows so that it cannot be interpreted easily. A strategy for avoiding the difficulty is constraining $\mathbf{\Lambda}$ to be sparse, i.e., it to have a number of zero elements. Among the sparse $\mathbf{\Lambda}$, the sparsest one is the most interpretable. Such $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p]' = (\lambda_{ij})$ can be formulated as

$$\boldsymbol{\lambda}_i \text{ being filled with zeros except a single element with unknown location.} \quad (1)$$

An FA procedure for obtaining $\mathbf{\Lambda}$ with (1) has been proposed by [3], in which a least squares approach is taken with the matrix-decomposition formulation of FA ([1], [7]). However, this procedure is not feasible for the data with $p > n$. On the other hand, a maximum likelihood (ML) FA procedure with EM-algorithm is feasible for such data, as shown in [2]. In this paper, we thus propose a modified MLFA procedure in which an EM algorithm is used for obtaining sparsest loadings. The modification is detailed in Section 3 after the EM algorithm for the standard FA is introduced in the next section.

## 2 EM Algorithm for FA

The normality of $\mathbf{f}$ and $\mathbf{e}$ is assumed as $\mathbf{f} \sim N_m(\mathbf{0}_m, \mathbf{\Phi})$ and $\mathbf{e} \sim N_p(\mathbf{0}_p, \mathbf{\Psi})$, with $\mathbf{0}_m$ the $m \times 1$ zero vector.

* Graduate Schools of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka 565-0871, Japan, E-mail: adachi@hus.osaka-u.ac.jp, Tel: +81-6-6879-4040.
† Department of Mathematics and Statistics, Open University, Walton Hall, Milton Keynes MK7 6AA, UK.

Using it in the FA model, we have $\mathbf{x} \sim N_p(\mathbf{0}_m, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}'+\mathbf{\Psi}$, which leads to the log-likelihood

$$f(\mathbf{\Lambda},\mathbf{\Psi},\mathbf{\Phi}) = -\log|\,\mathbf{\Lambda\Phi\Lambda}'+\mathbf{\Psi}| - \mathrm{tr}\mathbf{S}(\mathbf{\Lambda\Phi\Lambda}'+\mathbf{\Psi})^{-1}, \tag{2}$$

where $\mathbf{S}$ the $p \times p$ sample covariance matrix. For maximizing (2) over $\mathbf{\Lambda}$,$\mathbf{\Psi}$, and $\mathbf{\Phi}$, E- and M-steps are iterated in the EM-algorithm.

The E-step requires us to obtain

$$g(\mathbf{\Lambda},\mathbf{\Psi},\mathbf{\Phi}) = -\Sigma_j \log\psi_i - \mathrm{tr}(\mathbf{S}+\mathrm{tr}\mathbf{\Lambda Q\Lambda}'-2\mathrm{tr}\mathbf{C\Lambda}')\mathbf{\Psi}^{-1} - \log|\mathbf{\Phi}| - \mathrm{tr}\mathbf{Q\Phi}^{-1} \ . \tag{3}$$

Here, $\psi_i$ denotes the $j$th diagonal element of $\mathbf{\Psi}$, $\mathbf{Q} = \mathbf{A}'\mathbf{S}\mathbf{A} + \mathbf{U}$, and $\mathbf{C} = \mathbf{S}\mathbf{A}$, with $\mathbf{A}$ and $\mathbf{U}$ obtained as $\mathbf{A} = \mathbf{\Sigma}^{-1}\mathbf{\Lambda\Phi}$ and $\mathbf{U} = \mathbf{I}_m - \mathbf{A}'\mathbf{\Lambda\Psi}^{-1}\mathbf{\Phi}$ using $\mathbf{\Sigma}$, $\mathbf{\Lambda}$, and $\mathbf{\Phi}$ in the previous round of iteration.

The M-step requires to obtain $\mathbf{\Lambda}$,$\mathbf{\Psi}$, and $\mathbf{\Phi}$ that increase (3). The diagonal $\mathbf{\Psi}$ maximizing (3) is presented in [6]. For maximizing (3) over correlation matrix $\mathbf{\Phi}$, we can reparameterize it as $\mathbf{\Phi} = \mathbf{T}'\mathbf{T}$ with $\mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}_m$. It allows the task to be reformulate as the maximization as minimizing

$$f(\mathbf{T}) = \log|\mathbf{T}'\mathbf{T}| + \mathrm{tr}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{Q} + \mathit{Constant} \tag{4}$$

over $\mathbf{T}$ subject to $\mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}_m$ with *Constant* the term irrelevant to $\mathbf{T}$. The constrained minimization of (4) can be attained using the GP algorithm [5], in which $\mathbf{T}$ is iteratively updated as $\mathbf{T}_{\mathrm{new}} = \mathbf{V_T}\mathrm{diag}(\mathbf{V_T}\mathbf{V_T}')^{-1/2}$. Here, $\mathbf{V_T} = \mathbf{T} - \alpha\partial f(\mathbf{T})/\partial\mathbf{T}$ with $\alpha$ a positive value leading to $f(\mathbf{T}) \geq f(\mathbf{T}_{\mathrm{new}})$ and $\partial f(\mathbf{T})/\partial\mathbf{T} = 2\,|\,\mathbf{T}'\mathbf{T}\,|\,\mathbf{T}'^{-1} - 2\mathbf{T}'^{-1}\mathbf{Q}\mathbf{T}^{-1}\mathbf{T}'^{-1}$ . The remaining task is minimization of (3) over $\mathbf{\Lambda}$. As detailed in the next section, the modification with (1) is incorporated in the task.

## 3 Sparsest Modification

Let us consider maximizing (3) over $\mathbf{\Lambda}$ subject to (1). We can rewrite (3) as

$$f(\mathbf{T}) = const - \mathrm{tr}(\mathbf{\Lambda Q\Lambda}'-2\mathbf{C\Lambda}')\mathbf{\Psi}^{-1} = const - \textstyle\sum_{i=1}^{p} g_i(\lambda_i)/\psi_i \ , \tag{5}$$

with $\psi_i > 0$, *const* being an expression irrelevant to $\mathbf{\Lambda}$, and

$$g_i(\boldsymbol{\lambda}_i) = \boldsymbol{\lambda}_i'\mathbf{Q}\boldsymbol{\lambda}_i - 2\Sigma_j c_{ij}\lambda_{ij} = \Sigma_j\Sigma_k q_{jk}\lambda_{ij}\lambda_{ik} - 2\Sigma_j c_{ij}\lambda_{ij} \tag{6}$$

the function of the $i$th row of $\mathbf{\Lambda}$. It shows that the optimal $\mathbf{\Lambda}$ maximizing (5) subject to (1) can be obtained by minimizing (6) over $\boldsymbol{\lambda}_i$ under (1) for each $i$. Indeed, considering (1) and letting $J(i)$ denote the location of the element to be given a nonzero value in $\boldsymbol{\lambda}_i$, we can further rewrite $g_i(\boldsymbol{\lambda}_i)$ as

$$g_i(\boldsymbol{\lambda}_i) = q_{J(i),J(i)}\lambda_{i,J(i)}^2 - 2c_{i,J(i)}\lambda_{i,J(i)} = q_{J(i),J(i)}\left(\lambda_{i,J(i)} - \frac{c_{i,J(i)}}{q_{J(i),J(i)}}\right)^2 - \frac{c_{i,J(i)}^2}{q_{J(i),J(i)}} \geq -\frac{c_{i,J(i)}^2}{q_{J(i),J(i)}} \ . \tag{7}$$

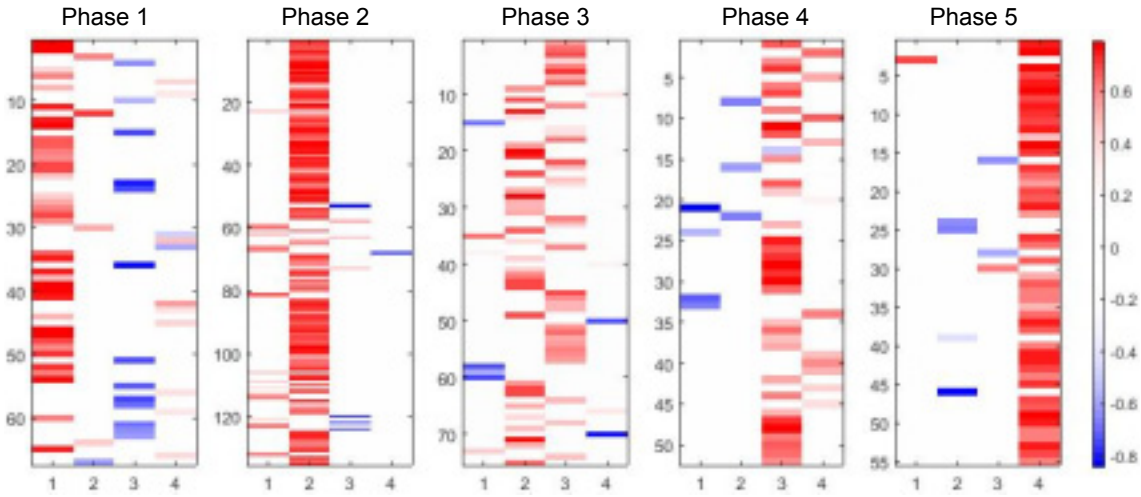
The inequality in (7), which follows from the positive-definiteness of $\mathbf{Q}$ implying $q_{jj} > 0$, shows that the lower limit of $g_i(\boldsymbol{\lambda}_i)$ is $-c_{i,J(i)}^2/q_{J(i),J(i)}$ which is attained for $\lambda_{i,J(i)} = -c_{i,J(i)}^2/q_{J(i),J(i)}$. Further, the optimal $J(i)$ is the index (from 1 to $m$) for which the limit $-c_{i,J(i)}^2/q_{J(i),J(i)}$ is minimal. This selection of the optimal nonzero loading can be expressed as

$$\lambda_{ij} = \begin{cases} c_{ij}^2/q_{jj} & \mathit{iff}\ \ j = \arg\min_{1\leq k\leq m} -c_{ik}^2/q_{kk} \\ 0 & \text{otherwise} \end{cases} \ . \tag{8}$$

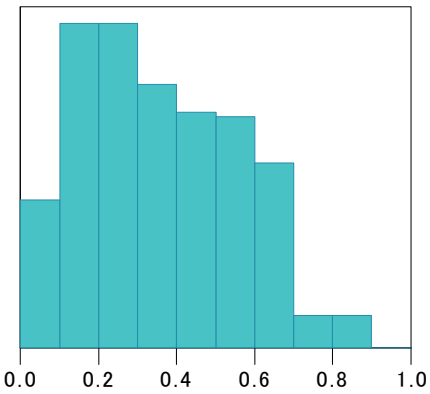
We can find that the formulas require in this and the last sections only require the nonnegative-definiteness of $\mathbf{S}$: it may not be positive-definite. Thus, the proposed method is feasible to data matrices with $p > n$.

## 4 Application

We performed the proposed method for the yeast cell cycle data matrix of $n = 17$ by $p = 384$ (genes)




(**A**) Heat maps of the genes × factors blocks in the loading matirx




(**B**) Histogram of unique variances

| Factor | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1 | -0.13 | 0.04 | 0.71 |
| 2 | | 1 | -0.52 | 0.08 |
| 3 | | | 1 | 0.06 |
| 4 | | | | 1 |

(**C**) Factor Correlations

Figure 1: Solution of the proposed factor analysis for gene expression data

presented by [8]. This data matrix, which is publicly available at http://faculty.washington.edu/kayee/pca, have been first log-transformed and then standardized so that the column averages and variances are zero and one, respectively. The solution resulting with $m$ set at 4 as in [4] is shown in Figure 1.

The solution of the loading matrix is presented block-wise in Figure 1(A). As the 384 genes are categorized into five phases of cell cycles [8], the five blocks (genes × components) in (A) correspond to the five phases. The loadings are considered to be reasonable, as each phase has a specific feature of loadings: [a] The genes in Phases 1, 2, and 4 are positively loaded by Factors 1, 2, and 3, respectively; [b] Phase 5 are characterized by positive loadings for Factor 4 and negative ones for 2; [c] Phase 3 consists of the genes positively loaded by Factor 2 or 3 and by both.

The histogram of the resulting unique variances is shown in Figure 1(B). Their range was [0.03, 0.89]: a solution with $\psi_j = 0$ was not found, which suggests that the unique variances were reasonably estimated.

Figure 1(C) shows the resulting factor correlations, where we can find factor 2 is negatively correlated with factor 3, while the factor 1 are considerably positively correlated with factor 4.

Those results are summarized as the path diagram in Figure 2.

## 5 Final Remarks

We proposed a sparsest FA procedure with an EM algorithm, which is feasible for high-dimensional

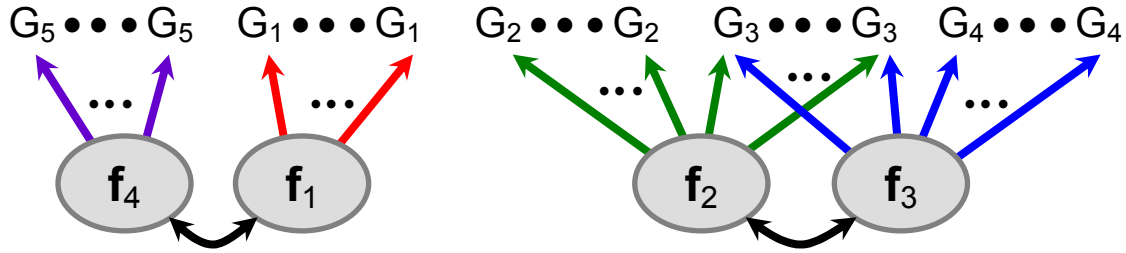



Figure 2: Path diagram for the solution in Figure 1 with $\mathbf{f}_\#$ and G standing for factors and genes, respectively.

data and provides a factor loading matrix with each row having single nonzero element. The application of the procedure to a gene expression data set demonstrated that genes can reasonably be classified into a few groups characterized by common factors.

**Acknowledgment**

This work is supported by a grant RPG-2013-211 from The Leverhulme Trust, UK and KAKENHI (26330039; Grant-in-Aid for Scientific Research (C)) from JSPS, Japan.

**References**

[1] Adachi, K.: Some contributions to data-fitting factor analysis with empirical comparisons to covariance-fitting factor analysis. *Journal of the Japanese Society of Computational Statistics*, 25(1): 25-38 (2012).

[2] Adachi, K. (2013). Factor analysis with EM algorithm never gives improper solutions when sample covariance and initial parameter matrices are proper. *Psychometrika*, 78(2): 380-394.

[3] Adachi, K. and Trendafilov, N.T. (2015). Sparse factor analysis for identifying optimal perfect simple structure. *Proceedings of the 2013 IASC satellite for the ISI WSC and the 8th IASC-ARS conference*: pp. 285-290.

[4] Adachi, K. and Trendafilov, N.T. (2015). Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics*, http://link.springer.com/article/10.1007/s00180-015-0608-4 (in press).

[5] Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika*, 67(7): 7-20.

[6] Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1): 69-76.

[7] Unkel, S., Trendafilov, N.T.: Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statisitical Review*, 78(3): 363-382 (2010).

[8] Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9): 763-774.