A Generalization of FA and PCA Which Includes Them as Special "*Sparse*" Cases

Kohei Adachi, Osaka University, Japan

1. PCA and FA as Matrix Factorization

Let $\mathbf{X}$ be an n-observations $\times$ p-variables column-centered data matrix, where the p variables are supposed to be explained by m common factors with rank($\mathbf{X}$) = $p > m$. By abbreviating "subject to" as "s.t.", principal component analysis (PCA) and factor analysis (FA) can be formulated as

$$\text{PCA: } \min_{\mathbf{F},\mathbf{A}} \|\mathbf{X} - \mathbf{FA}'\|^2 \text{ s.t. } n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m \text{ (the } m \times m \text{ identity matrix)}, \quad (1)$$

$$\text{FA: } \min_{\mathbf{F},\mathbf{A},\mathbf{U},\mathbf{\Psi}} \|\mathbf{X} - (\mathbf{FA}' + \mathbf{U\Psi})\|^2 = \|\mathbf{X} - [\mathbf{F},\mathbf{U}][\mathbf{A},\mathbf{\Psi}]'\|^2 \text{ s.t. } n^{-1}[\mathbf{F},\mathbf{U}]'[\mathbf{F},\mathbf{U}] = \mathbf{I}_{m+p}, \quad (2)$$

respectively[1,2]. Here, $\mathbf{F}$ ($n \times m$) is a matrix of PC/common-factor scores, $\mathbf{A}$ ($p \times m$) contains PC/factor loadings of variables, $\mathbf{\Psi}$ ($p \times p$) is diagonal, and the jth column of $\mathbf{U}$ ($n \times p$) uniquely affects that of $\mathbf{X}$. In a sparse version of PCA and FA, the constraint $cd(\mathbf{A}) = k$ or $cd(\mathbf{A}) \le k$ is added in (1) and (2)[3,4], with k a positive integer and $cd(\mathbf{A})$ the cardinality of $\mathbf{A}$ (i.e., its number of nonzero elements). This paper aims to present a formulation in which (1), (2), and their variants can be treated within a unified framework.

2. Generalized Sparse Component/Factor Analysis (GSCFA)

I propose a generalization of (1) and (2) which can include them and their sparse versions:

$$\min_{\mathbf{Z},\mathbf{C}} f(\mathbf{Z},\mathbf{C}) = \|\mathbf{X} - \mathbf{ZC}'\|^2 \text{ s.t. } n^{-1}\mathbf{Z}'\mathbf{Z} = \mathbf{I}_q \text{ and (3) } cd(\mathbf{C}) = k \text{ or } cd(\mathbf{C}) \le k \text{ for given } k \quad (3)$$

Here, $q \ge m + p$, $\mathbf{Z}$ is $n \times q$, and $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_q]$ ($p \times q$) contains the coefficients for $\mathbf{Z}$ with $cd(\mathbf{c}_j) \ge cd(\mathbf{c}_{j+1})$.

This problem may be called GSCFA, as it can give a PCA solution with $\mathbf{C} = [\mathbf{A}, \mathbf{O}]$ for $k = pm - m(m-1)/2$ and an FA solution with $\mathbf{C} = [\mathbf{A}, \mathbf{\Psi}, \mathbf{O}]$ for $k = pm - m(m-1)/2 + p$, with $\mathbf{O}$ a zero matrix. GSCFA (4) can also lead to a sparse PCA or FA solution (when k is a sufficiently small number). Further, GSCFA can provide a solution hybrid between PCA and FA, in which some variables are explained only by m common factors and others are affected by unique factors. Here, what type of solutions is given is unknown in advance.

3. Algorithm for Fixed Cardinality

GSCFA (4) can be solved by alternately iterating the two steps described in the next paragraphs.

In the first step, $f(\mathbf{Z},\mathbf{C})$ in (3) is minimized over $\mathbf{Z}$ s.t. $n^{-1}\mathbf{Z}'\mathbf{Z} = \mathbf{I}_q$ for given $\mathbf{C}$. It amounts to $\max_{\mathbf{Z}} g(\mathbf{Z}) = \text{tr}(\mathbf{XC})'\mathbf{Z}$ s.t. $n^{-1}\mathbf{Z}'\mathbf{Z} = \mathbf{I}_q$. The singular value decomposition of $\mathbf{XC}$ defined as $\mathbf{XC} = \mathbf{K\Lambda L}'$ leads to that $g(\mathbf{Z}) \le \text{tr}\mathbf{\Lambda}$ and the upper limit $\text{tr}\mathbf{\Lambda}$ is attained for $\mathbf{Z} = n^{1/2}(\mathbf{KL}' + \mathbf{K}_\perp\mathbf{L}_\perp')$ with $[\mathbf{K}, \mathbf{K}_\perp]'[\mathbf{K}, \mathbf{K}_\perp] = [\mathbf{L}, \mathbf{L}_\perp]'[\mathbf{L}, \mathbf{L}_\perp] = \mathbf{I}_q$[2].

In the next, $f(\mathbf{Z},\mathbf{C})$ is minimized over constrained $\mathbf{C}$, for given $\mathbf{Z}$. It should be noted that $f(\mathbf{Z},\mathbf{C}) = \|\mathbf{X} - \mathbf{ZS}_{XZ}\|^2 + n\|\mathbf{S}_{XZ} - \mathbf{C}\|^2$ under $n^{-1}\mathbf{Z}'\mathbf{Z} = \mathbf{I}_q$, with $\mathbf{S}_{XZ} = n^{-1}\mathbf{X}'\mathbf{Z}$. Thus, our task is $\min_{\mathbf{C}} \|\mathbf{S}_{XZ} - \mathbf{C}\|^2$ s.t. $cd(\mathbf{C}) = k$ / $cd(\mathbf{C}) \le k$, which can be attained without / with a penalty function, respectively[3,4].

The resulting $f(\mathbf{Z},\mathbf{C})$ value is expressed as $n\text{tr}\mathbf{S}_{XX}\{1 - PVE(k)\}$, where $\mathbf{S}_{XX} = n^{-1}\mathbf{X}'\mathbf{X}$ contains covariances and

$$PVE(k) = n^{-1}\|\mathbf{ZC}'\|^2/\text{tr}\mathbf{S}_{XX} = \text{tr}\mathbf{CC}'/\text{tr}\mathbf{S}_{XX} \quad (4)$$

is the *proportion* of the total *variance* in $\mathbf{X}$ *explained* by the model part $\mathbf{ZC}'$ for $f(\mathbf{Z},\mathbf{C}) = \|\mathbf{X} - \mathbf{ZC}'\|^2$ in (3).

4. Selection of Cardinality

For selecting a suitable k value in (3), $PVE(k)$ in (4) can be used. It increases monotonically with k, but (4) is normalized with its range [0, 1], which facilitates the choice of the lower limit PVE_L defining permissible (4) values $\ge PVE_L$. I thus consider the following steps for selecting a suitable k when PVE_L is given:

[S1] Perform PCA (1) for $\mathbf{X}$ with $m = 1$ (which implies $k = p$).

[S2] Set $k := k + 1$ and perform GSCFA (4).

[S3] Go back to [S2] if the resulting (4) value is lower than PVE_L; otherwise, accept the current solution.

Here, it should be noted that the number of (common and unique) factors is selected computationally.

References

[1] Adachi, K. (2016). *Matrix-based introduction to multivariate data analysis*. Springer.

[2] Adachi, K., & Trendafilov, N.T. (2018). Some mathematical properties of the matrix decomposition solution in factor analysis. *Psychometrika*, **83**, 407-424.

[3] Adachi, K., & Trendafilov, N.T. (2016). Sparse principal component analysis subject to prescribed cardinality of loadings. *Computational Statistics*, **31**, 1403-1427.

[4] Kawano, S., Matsui, H., & Hirose, K. (2018). *Statistical modeling via sparse estimation*. Kyoritsu-Shuppan (in Japanese).