Principal Component and Factor Analysis (unsupervised learning)

Marianna Bolla, Prof. DSc.

In multivariate data analysis, usually there are strong dependencies between the coordinates of the underlying random vector. With an appropriate transformation we want to describe its covariance structure by means of independent variables. In practical applications, there may be linear or near linear dependencies between the components of multidimensional data. To reduce the dimensionality, the first step is to simplify the covariance structure. If the underlying distribution is multivariate Gaussan (which is frequently the case due to the multidimensional Central Limit Theorem, after subtracting the means it suffices to treat the empirical covariance matrix of the observations, but in other situations (if the data is from a multivariate, absolutely continuous distribution) we can also be confined to the covariances. In such cases, instead of independence, we will speak of uncorrelatedness of the components.

Let $\mathbf{X}$ be a p-dimensional random vector with expectation $\mathbf{m}$ and covariance matrix $\mathbf{C}$. The *principal component transformation* assigns the following p-dimensional random vector $\mathbf{Y}$ to $\mathbf{X}$:

$$\mathbf{Y} = \mathbf{U}^T(\mathbf{X} - \mathbf{m}),$$

where $\mathbf{C} = \mathbf{U}\underline{\Lambda}\mathbf{U}^T$ is the spectral decomposition of the positive definite (possibly, semidefinite) covariance matrix $\mathbf{C}$. It is easy to see that the random vector $\mathbf{Y}$ has expectation $\mathbf{0}$ and covariance matrix $\underline{\Lambda}$. As $\underline{\Lambda}$ is a diagonal matrix, the components of $\mathbf{Y}$ are uncorrelated (in the Gaussian case also independent) with variances $\lambda_1 \ge \dots \ge \lambda_p \ge 0$, the diagonal entries of $\underline{\Lambda}$, i.e. the eigenvalues of $\mathbf{C}$. Denoting by $\mathbf{u}_1, \dots, \mathbf{u}_p$ the corresponding unit-norm eigenvectors, the ith component of $\mathbf{Y}$, called the ith *principal component* is

$$Y_i = \mathbf{u}_i^T(\mathbf{X} - \mathbf{m}), \quad i = 1, \dots, p.$$

In fact, Y_i is a linear combination of the components of $\mathbf{X}$ normalized such that $\|\mathbf{u}_i\| = 1$. The sum of the variances of the principal components is equal to the sum of the variances of X_i's, since

$$\sum_{i=1}^p \mathrm{Var}(Y_i) = \sum_{i=1}^p \lambda_i = \mathrm{tr}(\mathbf{C}) = \sum_{i=1}^p \mathrm{Var}(X_i)$$

and

$$\mathrm{Var}(Y_1) \ge \mathrm{Var}(Y_2) \ge \dots \ge \mathrm{Var}(Y_p) \ge 0.$$

Thus, we may say that the set of the principal components explains the total variation of the original random vector's components, in decreasing order. If $r = \mathrm{rank}(\mathbf{C}) < p$, then the principal components $Y_{r+1}, \dots, Y_p$ are zeros with probability 1.

By a linear algebra fact (see Proposition 4 of Lesson 1), the principal components can also be obtained as solutions of the following sequential maximization task:

$$\max_{\mathbf{v} \in \mathbb{R}^p,\, \|\mathbf{v}\|=1} \mathrm{Var}(\mathbf{v}^T(\mathbf{X} - \mathbf{m})) = \max_{\mathbf{v} \in \mathbb{R}^p,\, \|\mathbf{v}\|=1} \mathbf{v}^T\mathbf{C}\mathbf{v} = \lambda_1$$

and the maximum is attained with the choice $\mathbf{v} = \mathbf{u}_1$ (uniquely if $\lambda_1 > \lambda_2$). Hence, the first principal component $Y_1 = \mathbf{u}_1^T(\mathbf{X} - \mathbf{m})$ is obtained as the maximum variance, normalized linear combination of the components of $\mathbf{X}$. This was the $k = 1$ case. Further, for $k = 2, 3, \dots, r$, again in view of Proposition 4 of Lesson 1,

$$\max_{\substack{\mathbf{v}\in\mathbb{R}^p,\, \|\mathbf{v}\|=1 \\ \mathrm{Cov}(\mathbf{v}^T\mathbf{X}, Y_i)=0\ (i=1,\dots,k-1)}} \mathrm{Var}(\mathbf{v}^T(\mathbf{X} - \mathbf{m})) = \max_{\substack{\mathbf{v}\in\mathbb{R}^p,\, \|\mathbf{v}\|=1 \\ \mathbf{v}^T\mathbf{u}_i=0\ (i=1,\dots,k-1)}} \mathbf{v}^T\mathbf{C}\mathbf{v} = \lambda_k$$

and the maximum is attained with the choice $\mathbf{v} = \mathbf{u}_k$ (uniquely if $\lambda_k > \lambda_{k+1}$). Hence, the kth principal component $Y_k = \mathbf{u}_k^T(\mathbf{X} - \mathbf{m})$ is obtained as the maximum variance, normalized linear combination of the components of $\mathbf{X}$ under the condition of its uncorrelatedness with the preceding principal components $Y_1, \dots, Y_{k-1}$.

A more general statement is also true. For a fixed positive integer $k \le r$, such that $\lambda_k > \lambda_{k+1}$, the first k principal components provide the best rank k approximation of $\mathbf{X}$ in the following sense.

Proposition 1

$$\min_{\substack{\mathbf{A}\ is\, p\times p \\ \mathrm{rank}(\mathbf{A})=k}} \mathbb{E}\|\mathbf{X} - \mathbf{A}\mathbf{X}\|^2 = \mathbb{E}\|\mathbf{X} - \mathbf{P}\mathbf{X}\|^2,$$

where $\mathbf{P}$ *is the orthogonal projection onto* $\mathrm{Span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$.

Proof: Obviously,

$$\|\mathbf{X} - \mathbf{A}\mathbf{X}\|^2 \le \|\mathbf{X} - \mathbf{P}\mathbf{X}\|^2,$$

where $\mathbf{P}$ projects onto the subspace spanned by the coumn vectors of $\mathbf{A}$. Since $\mathbf{P}$ is a projection, $\mathbf{PX}$ and $\mathbf{X} - \mathbf{PX}$ are orthogonal. Therefore,

$$\mathbb{E}\|\mathbf{X}\|^2 = \mathbb{E}\|\mathbf{PX}\|^2 + \mathbb{E}\|\mathbf{X} - \mathbf{PX}\|^2,$$

where $\mathbb{E}\|\mathbf{X}\|^2$ is given, so minimizing $\mathbb{E}\|\mathbf{X} - \mathbf{PX}\|^2$ is equivalent to maximizing $\mathbb{E}\|\mathbf{PX}\|^2$.

Since $\mathbf{P} = \mathbf{B}\mathbf{B}^T$ with some suborthogonal matrix $\mathbf{B}$ ($\mathbf{B}^T\mathbf{B} = \mathbf{I}_k$),

$$\mathbb{E}\|\mathbf{PX}\|^2 = \mathbb{E}\mathrm{tr}(\mathbf{P}\mathbf{X}\mathbf{X}^T\mathbf{P}) = \mathrm{tr}\mathbf{PCP} = \mathrm{tr}\mathbf{B}\mathbf{B}^T\mathbf{C}\mathbf{B}\mathbf{B}^T =$$

$$= \mathrm{tr}\mathbf{B}^T\mathbf{C}\mathbf{B}\mathbf{B}^T\mathbf{B} = \mathrm{tr}\mathbf{B}^T\mathbf{C}\mathbf{B} = \sum_{i=1}^k \lambda_i(\mathbf{B}^T\mathbf{C}\mathbf{B}),$$

where $\lambda_i(\cdot)$ denotes the ith largest eigenvalue of the matrix in the argument.

By the Cauchy–Poincaré separation theorem it follows that $\lambda_i(\mathbf{B}^T\mathbf{C}\mathbf{B}) \le \lambda_i(\mathbf{C})$, $(i = 1, \dots, k)$. Equality is attained everywhere when $\mathbf{P}$ projects onto the subspace spanned by the eigenvectors corresponding to the k largest eigenvalues of $\mathbf{C}$, i.e. $\mathbf{P} = \sum_{i=1}^k \mathbf{u}_i\mathbf{u}_i^T$.

Note that

$$\mathbf{PX} = \sum_{i=1}^k \mathbf{u}_i\mathbf{u}_i^T\mathbf{X} = \sum_{i=1}^k \mathbf{u}_i Z_i,$$

which vector contains the first k principal components, otherwise zeros, in its coordinates (if the eigenvectors are the coordinate axes).

In fact, it is the ratio $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$ which tells us the proportion of $\mathbf{X}$'s total variation explained by the first k principal components. Therefore, it suffices to retain only the first k principal components if there is a remarkable gap in the spectrum of $\mathbf{C}$ between λ_k and λ_{k+1}. Based on a statistical sample $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$, where $n \gg p$, for $k = 0, \dots, r$ we test the hypothesis that the last $p-k$ eigenvalues of $\mathbf{C}$ are equal, until it is accepted. The likelihood ratio test statistic is based on the spectrum of the empirical covariance matrix. We perform a likelihood ratio test for testing the following sequence of null-hypotheses:

$$H_{0,k} : \lambda_{k+1} = \dots = \lambda_p \quad \text{for} \quad k = 0, 1, \dots, p-1$$

until accepted. By the asymptotic theory of the likelihood ratio tests, the transformed test statistic $-2 \ln T_{n,k}$ has the form

$$n(p-k) \ln \frac{a}{g} \quad \text{with} \quad a = \frac{\hat{\lambda}_{k+1} + \dots + \hat{\lambda}_p}{p-k}, \quad g = (\hat{\lambda}_{k+1} \dots \hat{\lambda}_p)^{\frac{1}{p-k}}$$

where $\hat{\lambda}_i$'s are the eigenvalues of the empirical covariance matrix $\hat{\mathbf{C}}$ of the sample, and a and g denote the algebraic and geometric means of the last $n-k$ eigenvalues of the empirical covariance matrix $\hat{\mathbf{C}}$, respectively. For "large" n, the test statistic asymptotically follows χ^2-distribution with degrees of freedom $\frac{1}{2}(p-k+2)(p-k-1)$, the decrease in the number of parameters under the assumption of $H_{0,k}$. Note that the number of eigenvalues (p) is decreased with $p-k-1$, whereas in the $p \times p$ orthogonal matrix (containing the eigenvectors), the number $(p-1)p/2$ of free parameters is decreased by $(p-k-1)(p-k)/2$, the number of free parametrs in a $(p-k) \times (p-k)$ rotation (in the eigensubspace corresponding to the multiple eigenvalue).

Given the significance, we stop if $H_{0,k}$ is accepted, which can be interpreted as the number of significant PC's is k. The PC's themselves are estimated from the sample via its mean vector and the spectral decomposition of $\hat{\mathbf{C}}$.

In the model of Factor Analysis, a smaller number of latent variables explain the correlations between the original ones. We say that the n-dimensional rv $\mathbf{X}$ has a factor structure if each variable X_i depends on a small number of latent common factors plus a component that is specific to X_i. Formally, $\mathbf{X}$ has a k-factor structure if it obeys the following model with the integer $1 \le k < p$:

$$\mathbf{X} = \mathbf{m} + \mathbf{B}\mathbf{f} + \mathbf{e} \tag{1}$$

where the components of the k-dimensional rv $\mathbf{f} = (f_1, \dots f_k)^T$ are the *common factors*, and the components of the p-dimensional rv $\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_p)^T$ are the *individual factors* (disturbances), whereas the $p \times k$ matrix $\mathbf{B} = (b_{ij})$ contains the *factor loadings.* We make the following assumptions:

$$\mathbb{E}(\mathbf{f}) = \mathbf{0}, \quad \text{Var}(\mathbf{f}) = \mathbf{I}_k, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{Var}(\mathbf{e}) = \mathbf{D}, \quad \text{Cov}(\mathbf{f}, \mathbf{e}) = \mathbf{O} \tag{2}$$

where $\mathbf{D}$ is a $p \times p$ diagonal matrix and the cross-covariance matrix of $\mathbf{f}$ and $\mathbf{e}$, denoted by $\text{Cov}(\mathbf{f}, \mathbf{e})$, is the $k \times n$ zero matrix. This means that both the common and the individual factors have uncorrelated components that are also uncorrelated with each other; further, the factors are normalized so that they have unit variances. If $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$, then $\mathbf{Y} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k)$ is a k-dimensional standard normal rv. However, its components cannot be obtained with an explicit transformation, like the PC's. The factors are latent variables that we cannot observe directly, we can only estimate the so-called *factor scores.*

To identify the model (1), consider the equation

$$\mathbf{C} = \mathbf{B}\mathbf{B}^T + \mathbf{D} \tag{3}$$

obtained by equating the covariance matrices. This equation is the basis for the ML-estimation of the rank k matrix $\mathbf{B}\mathbf{B}^T$ and the diagonal matrix $\mathbf{D}$; further, for testing the hypothesis that the number of factors is k. For the coordinates and variances of X_i's, Equations (1) and (3) provide

$$X_i = \mu_i + \sum_{j=1}^{k} b_{ij} f_j + e_i, \quad \mathrm{Var}(X_i) = \sum_{j=1}^{k} b_{ij}^2 + \mathrm{Var}(e_i), \quad i = 1, \dots, p.$$

That is, every X_i depends on all of the common factors f_j's, but only depends on its own individual factor e_i. Here $\sum_{j=1}^{k} b_{ij}^2$ is the part of the variance of X_i, accounted for the common factors, and it is called *communality* of X_i; this makes sense when, instead of $\mathbf{C}$, the correlation matrix of X_i's is used (it indeed has rational if X_i's are measured on different scales).

Via counting the number of parameters, it is proved that unique solution to (3) can be expected with the so-called Lederman bound

$$k \le \frac{1}{2}(2p + 1 - \sqrt{8p + 1}).$$

Also observe that the structure described in (1) and (2) is not sufficient to identify the factors and the factor loadings: if $\mathbf{Q}$ is a $k \times k$ orthogonal matrix, then $\mathbf{Q}\mathbf{f}$ and $\mathbf{B}\mathbf{Q}^{-1}$ fulfill (1) and (2) as well as $\mathbf{f}$ and $\mathbf{B}$ do. However, when the factors and factor loadings are linearly transformed as above, the common components $\sum_{j=1}^{k} b_{ij} f_j$ and the specific components e_i do not undergo any change. The selection of a particular vector of factors, that is, the *identification* of the factors, requires additional criteria. For example, one of the factors has no impact on some of the variables or the sum of the squares of the loadings of one of the factors is maximum. Such constraints also depend on the particular application. There is a great variety of FA methods, we consider the following two to be the most important:

- ML based FA: If we have an $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ distributed sample, then we maximize its log-likelihood function

  $$-\frac{1}{2} n \ln |\mathbf{C}| - \frac{1}{2} n \mathrm{tr} \mathbf{C}^{-1} \hat{\mathbf{C}} + \text{constant}$$

  with respect to $\mathbf{B}, \mathbf{D}$ subject to $\mathbf{C} = \mathbf{B}\mathbf{B}^T + \mathbf{D}$, where $|\mathbf{C}|$ is the determinant of $\mathbf{C}$ and $\hat{\mathbf{C}}$ is the sample covariance matrix, estimated from an independent, identically distributed (iid) sample. To avoid the ambiguity due to rotation, we also put the constraint that $\mathbf{B}^T \mathbf{D}^{-1} \mathbf{B}$ is diagonal. Equivalently, we have to solve

  $$\ln |\mathbf{B}\mathbf{B}^T + \mathbf{D}| + \mathrm{tr}(\mathbf{B}\mathbf{B}^T + \mathbf{D})^{-1} \hat{\mathbf{C}} \to \min, \text{ subject to } \mathbf{B}^T \mathbf{D}^{-1} \mathbf{B} \text{ diagonal.}$$

  There are both theoretical results and algorithms based on numerical methods at our disposal to treat this problem.

- PC based FA: If the variance of e_i does not depend on i, that is, $\mathbf{D} = \sigma^2 \mathbf{I}_p$ with some $\sigma > 0$, then the columns of $\mathbf{B}$ span the same linear space as the first k eigenvectors of $\mathbf{C}$ do. This is the rationale for using the first k principal components of $\mathbf{X}$ to estimate the factors and the factor loadings, a widespread though unwarranted practice. Actually, the Principal Component transformation yields $\mathbf{X} = \mathbf{m} + \mathbf{U}\mathbf{Y} = \mathbf{m} + (\mathbf{U}\underline{\Lambda}^{1/2})(\underline{\Lambda}^{-1/2}\mathbf{Y})$, that gives rise to estimate the factor loading matrix with $(\sqrt{\hat{\lambda}_1}\hat{\mathbf{u}}_1, \dots, \sqrt{\hat{\lambda}_k}\hat{\mathbf{u}}_k)$, where $\hat{\lambda}_i$'s and $\hat{\mathbf{u}}_i$'s are the first k eigenvalues and eigenvectors of $\hat{\mathbf{C}}$, and k is selected according to the spectral gap of $\hat{\mathbf{C}}$.

As an example from meteorology, suppose that X_i's are the yearly variations of average temperatures, observed in $p = 30$ European cities for $n = 60$ years (this is a sample). The factor structure above, with $k = 1$, would explain such a variation as depending on one common stochastic latent variable, plus local variables that have zero covariance with one another. However, FA was developed by psychometricians in the first half of the 20th century (Spearman, Thurstone), and was used to find latent common factors behind rv's corresponding to results of psychological tests. The very meaning of the factors, like general intelligence, was established by the experts, based on the loadings of the individual factors in the variables X_i's. The interpretation of the factors is the most straightforward if each variable is loaded highly on at most one factor, and if all the factor loadings are either large (in absolute value) or near zero, with few intermediate values. Then the variables can be divided into disjoint sets, each of which is associated with one factor, and some variables may be left over. The factor f_j can be interpreted as the common feature of those X_i's for which b_{ij} is large. We can make advantage of a $k \times k$ rotation $\mathbf{Q}$ such that the factor loading matrix $\mathbf{B}\mathbf{Q}^{-1}$ is the best interpretable in the above sense. For this convenience, there are methods of rotation elaborated, e.g., the VARIMAX rotation.