STA437/2005 - Methods for Multivariate Data

Gun Ho Jang

October 27, 2014

Principal Components

As the size of data gets larger, it is harder to handle and to conduct data analysis. A data reduction method is required. A *principal component analysis (PCA)* is one of the most popular for data reduction. It is concerned with explaining the variance-covariance matrix through a few linear combinations of variables. In other words, this method keeps the major pattern of the data structure but discards random noise.

Consider an i.i.d. random vectors $\mathbf{x}_i \in \mathbb{R}^p$ having variance Σ with very big p. Rather than working on $\mathbf{x}_i$'s directly, an appropriate linear transformation $\mathbf{y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ip})^\top = A\mathbf{x}_i$ so that $\mathbf{y}_{ij} = a_{j1}\mathbf{x}_{i1} + \cdots + a_{jp}\mathbf{x}_{ip}$ where $A = (a_{ij})_{p\times p}$ is invertible. A few more requirements are

- A must be chosen to make $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ip}$ uncorrelated, that is, $\mathrm{Cov}(\mathbf{y}_{ij}, \mathbf{y}_{ik}) = 0$ for any $j \neq k$.
- $\mathbb{V}ar(\mathbf{y}_{i1}) \geq \mathbb{V}ar(\mathbf{y}_{i2}) \geq \cdots \geq \mathbb{V}ar(\mathbf{y}_{ip})$

Such transformation can be obtained using eigen values and eigen vectors. Consider the spectral decomposition of $\Sigma = \mathbb{V}ar(\mathbf{x}_i)$, that is, for an orthonormal matrix U and a diagonal matrix Λ having non-increasing diagonals, $\Sigma = U\Lambda U^\top$ with $\Lambda_{11} \geq \Lambda_{22} \geq \cdots \Lambda_{pp}$. Then take $A = U^\top$ so that $\mathbf{y}_i = U^\top\mathbf{x}_i$

Proposition. Suppose $\mathbf{x}_i$'s are i.i.d. with variance Σ which having spectral decomposition $\Sigma = U\Lambda U^\top$ with $\Lambda_{11} \geq \Lambda_{22} \geq \cdot \geq \Lambda_{pp}$. Then $\sum_{j=1}^p \mathbb{V}ar(\mathbf{x}_{ij}) \sum_{j=1}^p \mathbb{V}ar(\mathbf{y}_{ij}) = \sum_{j=1}^p \Lambda_{jj}$.

Proof. Note that $\sum_{j=1}^p \mathbb{V}ar(\mathbf{x}_{ij}) = \sum_{j=1}^p \Sigma_{jj} = \mathrm{tr}(\Sigma) = \mathrm{tr}(U\Lambda U^\top) = \mathrm{tr}(U^\top U\Lambda) = \mathrm{tr}(\Lambda) = \sum_{j=1}^p \Lambda_{jj} = \sum_{j=1}^p \mathbb{V}ar(\mathbf{y}_{ij})$. □

If the variance of $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ik}$ explains the most variance of $\mathbf{x}_i$, then it is enough to analyze $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ik})$ rather than $\mathbf{y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ip})^\top$ or equivalently $\mathbf{x}_i$. If more than 80% or 90% of variance can be attributed to the first a few PCs, say k components, then the large p vectors can be shrunken down by the first k PCs.

If there is a random variable having extremely large variance, then the first principal component is dominated by such random variables. If this phenomenon happened due to the measurement scale, PCA results is spurious and unreliable. In such cases, a PCA on standardized data is recommended which is equivalent to the PCA on correlation.

Two special cases

Case I: Uncorrelated random variables.
Principal components are random variables having large variance.

Case II: Equal correlation random variables.
Consider $\Sigma = (\Sigma_{ij})$ with $\Sigma_{ij} = \rho(\Sigma_{ii}\Sigma_{jj})^{1/2}$ for any $i \neq j$. The correlation matrix becomes

$$R = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = (1-\rho)I_p + \rho\mathbf{1}_p\mathbf{1}_p^\top.$$

The largest eigen value is $\lambda_1 = 1 + (p-1)\rho$ with associated eigen vector $\mathbf{u}_1 = \mathbf{1}/\sqrt{p}$. The remaining $p-1$ eigen values are $\lambda_2 = \cdots = \lambda_p = 1-\rho$. Hence the first principal component explains $\lambda_1/(\lambda_1 + \cdots + \lambda_p) = (1+(p-1)\rho)/(1+(p-1)\rho + (p-1)(1-\rho)) = (1+(p-1)\rho)/p = \rho + (1-\rho)/p$.

Example. If the variance is $\Sigma = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$, the the eigen values and vectors are $\boldsymbol{\lambda} = (100.161, 0.837)$ and $\mathbf{u}_1 = (0.040, 0.999)^\top$, $\mathbf{u}_2 = (-0.999, 0.040)^\top$. Hence, X_2 contributes the most part of the first principal component.

Suppose X_2 is originally measured in millimeters. If it is measured in centimeters, then $\widetilde{\Sigma} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$ has eigen values and vectors $\widetilde{\boldsymbol{\lambda}} = (1.4, 0.6)^\top$, $\widetilde{\mathbf{u}}_1 = (0.707, 0.707)^\top$, $\widetilde{\mathbf{u}}_2 = (-0.707, 0.707)^\top$.

Even further, if it is measured in meters, then $\Sigma^\dagger = \begin{pmatrix} 1 & 0.004 \\ 0.004 & 0.0001 \end{pmatrix}$ and its eigen values and vectors are $\boldsymbol{\lambda}^\dagger = (1, 8.4 \times 10^{-5})^\top$ and $\mathbf{u}_1^\dagger = (-1, -0.004)^\top$, $\mathbf{u}_2^\dagger = (0.004, -1)^\top$.

Scale difference makes different interpretation of principal components.

Hence PCA on correlation is recommended if the scales of random variables are unintentionally varying.

In statistical genetics, batch effects were detected through principal component analyses.

The first principal component is a linear combination $\mathbf{a}_1^\top \mathbf{x}_j$ which maximizes the sample variance of $\mathbf{a}_1^\top \mathbf{x}_j$ subject to $\mathbf{a}_1^\top \mathbf{a}_1 = 1$. Then the ith principal component for $i > 1$ is the linear combination $\mathbf{a}_i^\top \mathbf{x}_j$ maximizing the sample variance subject to $\mathbf{a}_i^\top \mathbf{a}_i = 1$ and to make $(\mathbf{a}_1^\top x_j, \ldots, \mathbf{a}_i^\top \mathbf{x}_j)$ (pairwise) uncorrelated.

Example (Socioeconomic Variables)**.** Consider five socioeconomic variables: total population (in thousands), professional degree (percent), employed age over 16 (percentage), government employment (percentage), median home value (in hundred thousand dollars).

$$\bar{\mathbf{x}}^\top = \begin{pmatrix} 4.469 & 3.962 & 71.420 & 26.915 & 1.636 \end{pmatrix}$$

$$S = \begin{pmatrix} 3.397 & -1.102 & 4.306 & -2.078 & 0.027 \\ 1.102 & 9.673 & -1.513 & 10.953 & 1.203 \\ 4.306 & -1.513 & 55.626 & -28.937 & -0.044 \\ -2.078 & 10.953 & -28.937 & 89.067 & 0.957 \\ 0.027 & 1.203 & -0.044 & 0.957 & 0.319 \end{pmatrix}$$

The PC coefficients are

Variable	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$
X_1	0.039	-0.105	0.492	-0.863	-0.009
X_2	-0.071	-0.13	-0.864	-0.48	-0.015
X_3	-0.188	0.961	-0.046	-0.153	0.125
X_4	0.977	0.171	-0.091	-0.03	0.082
X_5	-0.058	-0.139	0.005	0.007	0.989
$\hat{\lambda}_i$	107.015	39.672	8.371	2.868	0.155
Cumulative percentage	0.677	0.928	0.981	0.999	1

The PC coefficients on correlations are

Variable	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$
X_1	0.263	-0.593	0.326	-0.479	-0.493
X_2	0.463	0.326	0.605	-0.252	0.5
X_3	0.784	-0.164	-0.225	0.551	-0.069
X_4	-0.217	0.145	0.663	0.572	-0.407
X_5	0.235	0.703	-0.194	-0.277	-0.58
$\hat{\lambda}_i$	1.992	1.368	0.864	0.535	0.241
Cumulative percentage	0.398	0.672	0.845	0.952	1

Large Sample Property

Proposition. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. $N_p(\mu, \Sigma)$ and $S = \widehat{U}\widehat{\Lambda}\widehat{U}^\top$. Then
(a) $\sqrt{n}(\widehat{\Lambda} - \Lambda)\mathbf{1}_p \approx N_p(O, 2\Lambda^2)$
(b) $\sqrt{n}(\widehat{\mathbf{u}}_i - \mathbf{u}_i) \approx N_p(O, \mathbf{U}_i)$ where

$$\mathbf{U}_i = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{u}_k \mathbf{u}_k^\top.$$

Consequently, a γ-confidence interval for λ_i can be obtained by

$$\frac{\widehat{\lambda}_i}{1 + z_{(1+\gamma)/2}\sqrt{2/n}} \leq \lambda_i \leq \frac{\widehat{\lambda}_i}{1 - z_{(1+\gamma)/2}\sqrt{2/n}}.$$

Testing for equal correlation

The hypothesis of interest is $H_0 : \boldsymbol{\rho} = (1-\rho)I_p + \rho\mathbf{1}\mathbf{1}^\top$.
Let $R = (\mathrm{cor}(\mathbf{x}_{1i}, \mathbf{x}_{1j}))$. Define $\bar{r}_k = \frac{1}{p-1}\sum_{i=1}^p r_{ik}$, $\bar{r} = \frac{2}{p(p-1)}\sum_{i<k} r_{ik}$ and $\hat{\gamma} = \frac{(p-1)^2(1-(1-\bar{r})^2)}{p-(p-2)(1-\bar{r})^2}$. Then

$$T = \frac{n-1}{(1-\bar{r})^2}\Big(\sum_{i<k}(r_{ik} - \bar{r})^2 - \hat{\gamma}\sum_{k=1}^p (\bar{r}_k - \bar{r})^2 \approx \chi^2((p+1)(p-2)/2).$$

Example. A genetic example is considered with $n = 150$,

$$\bar{\mathbf{x}} = \begin{pmatrix} 39.88 \\ 45.08 \\ 48.11 \\ 49.95 \end{pmatrix} \text{ and } R = \begin{pmatrix} 1 & 0.7501 & 0.6329 & 0.6363 \\ 0.7501 & 1 & 0.6925 & 0.7386 \\ 0.6329 & 0.6925 & 1 & 0.6625 \\ 0.6363 & 0.7386 & 0.6625 & 1 \end{pmatrix}$$

Its eigen values and vectors are

$$\boldsymbol{\lambda} = \begin{pmatrix} 3.058 \\ 0.382 \\ 0.342 \\ 0.217 \end{pmatrix} \text{ and } \mathbf{U} = \begin{pmatrix} -0.494 & -0.522 & -0.487 & -0.497 \\ 0.713 & 0.191 & -0.585 & -0.335 \\ -0.233 & 0.143 & -0.645 & 0.714 \\ 0.44 & -0.819 & 0.061 & 0.363 \end{pmatrix}$$

Then $(\bar{r}_j) = (0.6731, 0.7271, 0.6626, 0.6791)^\top$, $\bar{r} = 0.685$ and $\hat{\gamma} = 0.6855$. Hence,

$$T = \frac{n-1}{(1-\bar{r})^2}\Big(\sum_{i<k}(r_{ik} - \bar{r})^2 - \hat{\gamma}\sum_{k=1}^p (\bar{r}_k - \bar{r})^2 = \frac{149}{0.989}(0.01276 - 2.1329 \times 0.002445) = 11.362 > 11.071 = \chi^2_{0.95}(5) = \chi^2_\gamma((p+$$

Hence the hypothesis H_0 is rejected at the significance level 5% but it is not much strong.