Principal Components – I

• When a very large number $p$ of variables is measured on each sample unit, interpreting results of analyses might be difficult.

• It is often possible to reduce the dimensionality of the data by finding a smaller set $k$ of linear combinations of the original $p$ variables that preserve most of the variability across sample units.

• These linear combinations are called principal components.

• Finding the principal components (or PCs) is often an early step in a complex analysis. Scores for sample PCs can be used as response variables to fit MANOVA or regression models, to cluster sample units or to build classification rules.

• Principal components can also be used in diagnostic investigations to find outliers or high leverage cases.
Principal Components - I

- Geometrically, PCs are a new coordinate system obtained by rotating the original system which has $X_1, X_2, ..., X_p$ as the coordinate axes.

- The new set of axes represent the directions with maximum variability and permit a more parsimonious description of the covariance matrix (essentially by ignoring coordinates with little variability).

- PCs are derived from the eigenvectors of a covariance matrix $\Sigma$ or from the correlation matrix $\rho$, and do not require the assumption of multivariate normality.

- If multivariate normality holds, however, some distributional properties of PCs can be established.
Population Principal Components

• Let the random vector $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ have covariance matrix $\Sigma$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

• Consider $p$ linear combinations

$$
Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p
$$
$$
Y_2 = a_2'X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p
$$
$$
\vdots 
$$
$$
Y_p = a_p'X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
$$

• We have that

$$
\text{Var}(Y_i) = a_i'\Sigma a_i, \quad i = 1, 2, \ldots, p
$$
$$
\text{Cov}(Y_i, Y_k) = a_i'\Sigma a_k, \quad i, k = 1, 2, \ldots, p.
$$
Population Principal Components

• Principal Components are uncorrelated linear combinations $Y_1, Y_2, ..., Y_p$ determined sequentially, as follows:
  - The first PC is the linear combination $Y_1 = a'_1 X$ that maximizes $\text{Var}(Y_1) = a'_1 \Sigma a_1$ subject to $a'_1 a_1 = 1$.
  - The second PC is the linear combination $Y_2 = a'_2 X$ that maximizes $\text{Var}(Y_2) = a'_2 \Sigma a_2$ subject to $a'_2 a_2 = 1$ and $\text{Cov}(Y_1, Y_2) = a'_1 \Sigma a_2 = 0$.
    
    $\vdots$

  - The $i$th PC is the linear combination $Y_i = a'_i X$ that maximizes $\text{Var}(Y_i) = a'_i \Sigma a_i$ subject to $a'_i a_i = 1$ and $\text{Cov}(Y_i, Y_k) = a'_i \Sigma a_k = 0$ for $k < i$.
    
    $\vdots$
Population Principal Components

• Let $\Sigma$ have eigenvalue-eigenvector pairs $(\lambda_i, e_i)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then, the $i$th principal component is given by

$$Y_i = e_i'X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \quad i = 1, \ldots, p.$$ 

• Then,

$$\text{Var}(Y_i) = e_i'\Sigma e_i = e_i'\lambda_i e_i = \lambda_i, \quad \text{since } e_i' e_i = 1$$

$$\text{Cov}(Y_i, Y_k) = e_i'\Sigma e_k = e_i'\lambda_k e_k = 0, \quad \text{since } e_i' e_k = 0$$

• Proof: See section 8.2 in the textbook. Proof uses the maximization of quadratic forms that we studied in Chapter 2, with $B = \Sigma$. 

577
Population Principal Components

• Using properties of the trace of $\Sigma$ we have

\[
\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_i \text{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_i \text{Var}(Y_i).
\]

• Proof: We know that $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \text{tr}(\Sigma)$, and we also know that since $\Sigma$ is positive definite, we can decompose $\Sigma = P\Lambda P'$, where $\Lambda$ is the diagonal matrix of eigenvalues and $P$ is orthonormal with columns equal to the eigenvectors of $\Sigma$. Then

\[
\text{tr}(\Sigma) = \text{tr}(P\Lambda P') = \text{tr}(\Lambda P'P) = \text{tr}(\Lambda) = \lambda_1 + \lambda_2 + \cdots + \lambda_p.
\]
Population Principal Components

- Since the total population variance, $\sum_i \sigma_{ii}$, is equal to the sum of the variances of the principal components, $\sum_i \lambda_i$, we say that

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

is the proportion of the total variance associated with (or explained by) the $k$th principal component.

- If a large proportion of the total population variance (say 80% or 90%) is explained by the first $k$ PCs, then we can ignore the original $p$ variables and restrict attention to the first $k$ PCs without much loss of information.
Population Principal Components

- The correlation between the $i$th PC and the $k$th original variable,

$$\rho_{Y_i, X_k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

is a measure of the contribution of the $k$th variable to the $i$th principal component.

- An alternative is to consider the coefficients $e_{ik}$ to decide whether variable $X_k$ contributes greatly or not to $Y_i$.

- Ranking the $p$ original variables $X_k$ with respect to their contribution to $Y_i$ using the correlations or the coefficients $e_{ik}$ results in similar conclusions when the $X_k$ have similar variances.
PCs from Multivariate Normal Populations

- Suppose that $X \sim \mathcal{N}_p(\mu, \Sigma)$. The density of $X$ is constant on ellipsoids centered at $\mu$ that are determined by

$$ (x - \mu)'\Sigma^{-1}(x - \mu) = c^2, $$

with axes $\pm c\sqrt{\lambda_i}e_i$, $i = 1, \ldots, p$, with $(\lambda_i, e_i)$ the eigenvalue-eigenvector pairs of $\Sigma$.

- Assume for simplicity that $\mu = 0$, so that

$$ c^2 = x'\Sigma^{-1}x = \frac{1}{\lambda_1}(e'_1x)^2 + \frac{1}{\lambda_2}(e'_2x)^2 + \cdots + \frac{1}{\lambda_p}(e'_p x)^2, $$

using the decomposition of $\Sigma^{-1}$. Notice that $e'_ix$ is the $i$th PC of $x$.  

581
PCs from Multivariate Normal Populations

- If we let $y_i = e_i'x$ we can write
  
  $$c^2 = \frac{1}{\lambda_1}y_1^2 + \frac{1}{\lambda_2}y_2^2 + \cdots + \frac{1}{\lambda_p}y_p^2,$$

  which also defines an ellipse where the coordinate system has axes $y_1, y_2, \ldots, y_p$ lying in the directions of the eigenvectors.

- Thus, $y_1, y_2, \ldots, y_p$ are parallel to the directions of the axes of a constant density ellipsoid.

- Any point on the $i$th ellipsoid axis has $x$ coordinates proportional to $e_i'$ and has principal component coordinates of the form $[0, 0, \ldots, y_i, \ldots 0]$.

- If $\mu \neq 0$ then the mean-centered PC $y_i = e_i'(x - \mu)$ has mean 0 and lies in the direction of $e_i$. 

582
Two examples of PCs from MVN data

(a) $\hat{\lambda}_1 > \hat{\lambda}_2$

(b) $\hat{\lambda}_1 = \hat{\lambda}_2$

$$(x - \bar{x})' S^{-1} (x - \bar{x}) = c^2$$
**PCs from Standardized Variables**

- When variables are measured on different scales is useful to standardize the variables before extracting the PCs. If

\[ Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{ii}}} , \]

then \( Z = (V^{1/2})^{-1}(X - \mu) \) is the \( p \)-dimensional vector of standardized variables. The matrix \( V^{1/2} \) is diagonal with \( \sqrt{\sigma_{ii}} \) elements.

- Note that \( \text{Cov}(Z) = Corr(X) \), the correlation matrix of the original variables \( X \).

- If now \((\lambda_i, e_i)\) denote the eigenvalue-eigenvector pairs of \( Corr(X) \), the PCs of \( Z \) are given by

\[ Y_i = e_i'Z = e_i'(V^{1/2})^{-1}(X - \mu), \quad i = 1, \ldots, p. \]
PCs from Standardized Variables

- Proceeding as before,

\[
\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p
\]

\[
\rho_{Y_i,Z_k} = e_{ik} \sqrt{\lambda_i}, \quad i, k = 1, \ldots, p.
\]

Further,

\[
\begin{pmatrix}
\text{Proportion of standardized population variance due to } k\text{th principal component}
\end{pmatrix} = \frac{\lambda_k}{p}, \quad k = 1, \ldots, p,
\]

where the \( \lambda_i \) are the eigenvalues of \( \rho \).
PCs from Standardized Variables

- In general, the PCs extracted from $\Sigma=\text{Cov}(X)$ and from $\text{Corr}(X)$ will not be the same.

- The two sets of PCs are not even functions of one another, so standardizing variables has important consequences.

- When should one standardize before computing PCs?
PCs from Standardized Variables

- If variables are measured on very different scales (e.g. patient weights in kg vary from 40 to 100, protein concentration in ppm varying between 1 and 10), then the variables with the larger variances will dominate.

- When one variable has a much larger variance than any of the other variables, we will end up with a single PC that is essentially proportional to the dominating variable. In some cases, this might be desirable, but in several cases this may not be, in which case it would make sense to use the correlation matrices.
PCs from Uncorrelated Variables

• If variables $X$ are uncorrelated, then $\Sigma$ is diagonal with elements $\sigma_{ii}$.

• The eigenvalues in this case are $\lambda_i = \sigma_{ii}$, if the $\sigma_{ii}$s are in decreasing order and one choice for the corresponding eigenvectors is $e_i = \begin{bmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \end{bmatrix}'$.

• Since $e_i'X = X_i$ we note that the PCs are just the original variables. Thus, we gain nothing by trying to extract the PCs when the $X_i$ are uncorrelated.

• If the uncorrelated variables are standardized, the resulting PCs are just the original standardized variables $Z_i$.

• The above argument changes somewhat if the variances in $\Sigma$ are not in decreasing order: in that case, $e_i$’s together form an appropriate permutation matrix.
Sample Principal Components

- If \( x_1, x_2, ..., x_n \) is a sample of \( p \)-dimensional vectors from a distribution with \( (\mu, \Sigma) \), then the sample mean vector, sample covariance matrix and sample correlation matrix are \( \bar{x}, S, R \), respectively.

- Eigenvalue-eigenvector pairs of \( S \) are denoted \( (\hat{\lambda}_i, \hat{e}_i) \) and the \( i \)th sample PC is given by
  \[
  \hat{y}_i = \hat{e}'_i x = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \cdots + \hat{e}_{ip} x_p, \quad i = 1, ..., p.
  \]

- The sample variance of the \( i \)th PC is \( \hat{\lambda}_i \) and the sample covariance between \( (\hat{y}_i, \hat{y}_k) \) is zero for all \( i \neq k \).

- The total sample variance \( s_{11} + s_{22} + ... + s_{pp} \) is equal to \( \hat{\lambda}_1 + ... + \hat{\lambda}_p \) and the relative contribution of the \( k \)th variable to the \( i \)th sample PC is given by \( r_{\hat{y}_i, x_k} \).
Sample Principal Components

- Observations are sometimes centered by subtracting the mean: $(x_j - \bar{x})$.

- The PCs in this case are $\hat{y}_i = \hat{e}_i(x - \bar{x})$ where again $\hat{e}_i$ is the $i$th eigenvector of $S$.

- If $\hat{y}_{ji}$ is the score on the $i$th sample PC for the $j$th sample unit, then we note that the sample mean of the scores for the $i$th sample PC (across sample units) is zero:

$$\bar{\hat{y}}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{e}_i'(x_j - \bar{x}) = \frac{1}{n} \hat{e}_i' \sum_{j=1}^{n} (x_j - \bar{x}) = \frac{1}{n} \hat{e}_i' 0 = 0.$$ 

- The sample variance of the $i$th sample PC is $\hat{\lambda}_i$, the $i$th largest eigenvalue of $S$. 

590
**PCs from the MLE of \( \Sigma \)**

- If data are normally distributed, \( \widehat{\Sigma} = \frac{1}{n} \sum_{j=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' \) is the MLE of \( \Sigma \), and for \((\widehat{\delta}, \widehat{e})\), the eigenvalue-eigenvector pairs of \( \widehat{\Sigma} \), the PCs
  \[
  \widehat{y}_{ji} = \widehat{e}_i x_j, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n
  \]
  are the MLEs of the population principal components.

- Also the eigenvalues of \( \widehat{\Sigma} \) are
  \[
  \widehat{\delta}_i = \frac{n - 1}{n} \lambda_i,
  \]
  the eigenvalues of \( S \) and the eigenvectors are the same. Then, both \( S \) and \( \widehat{\Sigma} \) give the same PCs with the same proportion of total variance explained by each.