• Sample Principal Components

Suppose the data $x_1, x_2, \ldots, x_n$ represent $n$ independent drawings from some $p$-dimensional population with mean vector $\mu$ and covariance matrix $\Sigma$. These data yield the sample mean vector $\bar{x}$, the sample covariance matrix $S$, and the sample correlation matrix $R$.

Motivation: construct uncorrelated linear combinations of the measured characteristics that account for much of the variation in the sample

Finding principal components

- replace population distribution by empirical distribution
- replace $\Sigma$ by $S$ (or $S_n$)
- replay $\rho$ by $R$
- then, the rest is the same as above

PCA based on $S$

If $S = \{s_{ik}\}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\lambda_1, \hat{e}_1), (\lambda_2, \hat{e}_2), \ldots, (\lambda_p, \hat{e}_p)$, the $i$th sample principal component is given by

$$\hat{y}_i = \hat{e}_i^T x = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \cdots + \hat{e}_{ip} x_p, \quad i = 1, 2, \ldots, p$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $x$ is any observation on the variables $X_1, X_2, \ldots, X_p$. Also,

Sample variance($\hat{y}_k$) = $\lambda_k, \quad k = 1, 2, \ldots, p$

Sample covariance($\hat{y}_i, \hat{y}_k$) = 0, \quad $i \neq k$

In addition, $\sum_i s_{ii} = \lambda_1 + \lambda_2 + \cdots + \lambda_p$

and

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik} \sqrt{\lambda_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p$$

- principal component scores

$\hat{y}_i = \hat{e}_i^T x, \quad i = 1, 2, \ldots, p$

$\hat{y}_i = \hat{e}_i^T (x - \bar{x}), \quad i = 1, 2, \ldots, p$

PCA based on $R$

If $z_1, z_2, \ldots, z_n$ are standardized observations with covariance matrix $R$, the $i$th sample principal component is

$$\hat{y}_i = \hat{e}_i^T z = \hat{e}_{i1} z_1 + \hat{e}_{i2} z_2 + \cdots + \hat{e}_{ip} z_p, \quad i = 1, 2, \ldots, p$$

where $(\hat{\lambda}_i, \hat{e}_i)$ is the $i$th eigenvalue-eigenvector pair of $R$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Also,

Sample variance ($\hat{y}_i$) = $\hat{\lambda}_i, \quad i = 1, 2, \ldots, p$

Sample covariance ($\hat{y}_i, \hat{y}_k$) = 0, \quad $i \neq k$

In addition, $\text{tr}(R) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$

and

$$r_{\hat{y}_i, z_k} = \hat{e}_{ik} \sqrt{\hat{\lambda}_i}, \quad i, k = 1, 2, \ldots, p$$

- principal component scores

$\hat{y}_i = \hat{e}_i^T z$
Number of Principal Components
- **Q**: how many principal components to retain?
- scree plot
- amount of total variance explained (say, 70~90%)
- for PCA based on $\rho$, use the cutoff point 1 or 0.7
  (i.e., keep PCs with $\hat{\lambda}_i \geq 1$ or 0.7)

Using Principal Components to display multivariate data
- In addition to plotting $X_1, \ldots, X_p$, plot $Y_1, \ldots, Y_k$ to check normality assumption and detect suspect observations
- The last few PCs can help pinpoint suspect observations

Large Sample Inference
- **Q**: What are the distributions of the $(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$, $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_p)$, $(\hat{y}_1, \ldots, \hat{y}_p)$?
- assumption
  - $X_1, X_2, \ldots, X_n$ are a random sample from a normal population.
  - eigenvalues of $\Sigma$ are distinct and positive, so that $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$.
- asymptotic distribution
  1. Let $\Lambda$ be the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\Sigma$, then $\sqrt{n} (\hat{\Lambda} - \Lambda)$ is approximately $N_p(0, 2\Lambda^2)$.

2. Let
   \[
   E_i = \lambda_i \sum_{k=1}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k'
   \]
   then $\sqrt{n} (\hat{e}_i - e_i)$ is approximately $N_p(0, E_i)$.

3. Each $\hat{\lambda}_i$ is distributed independently of the elements of the associated $\hat{e}_i$.
   - A large sample 100(1 - $\alpha$)% confidence interval for $\lambda_i$ is thus provided by
     \[
     \frac{\hat{\lambda}_i}{(1 + z(\alpha/2)\sqrt{2/n})} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{(1 - z(\alpha/2)\sqrt{2/n})}
     \]
   - Result 2 implies that the $\hat{e}_i$‘s are normally distributed about the corresponding $e_i$’s for large samples. The elements of each $\hat{e}_i$ are correlated, and the correlation depends to a large extent on the separation of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$.

Some important issues
- Analyses of principal components are more of a means to an end rather than an end in themselves, because they frequently serve as intermediate steps in much larger investigation.
- PCs with variance almost zero (i.e., $\hat{\lambda}_i \approx 0$)
  \[
  \hat{y}_i = \hat{\epsilon}_i'x = \hat{\epsilon}_{i1}x_1 + \hat{\epsilon}_{i2}x_2 + \cdots + \hat{\epsilon}_{ip}x_p \approx c, \ c : \text{a constant}
  \]
  ⇒ an unusually small value for eigenvalue can indicate a linear dependency in the data set
equal eigenvalues (i.e., $\lambda_i = \lambda_{i+1} = \cdots = \lambda_j$)

- population principal component
- sample principal component

- selecting a subset of variables $X_1, \ldots, X_p$

PCA does not quite reduce the data since PCs are linear combinations of all variables $X_1, \ldots, X_p$ so all variables are in some sense still needed. The technique, however, can help us discard some variables by keeping only those with the highest factor loading

- connection to the singular value decomposition (SVD)

Let us apply a SVD to the matrix $X - \bar{X}'$

$$(X - \bar{X}')_{n \times p} = U_{n \times p}L_{p \times p}V'_{p \times p}$$

where $U'U = I_m$, $V'V = I_m$, and $L$ is a diagonal matrix. Then,

$$(n - 1)S = (X - \bar{X}')'(X - \bar{X}') = VLU'ULV' = V(LL)V' = VL^2V'$$

$S = V\left(\frac{1}{n-1}L^2\right)V' = V\left(\frac{1}{\sqrt{n-1}}L\right)^2V' \equiv VL^2V'$$

which means

- columns of $V$: eigenvectors of $S$
- $\frac{l_i^2}{n-1}$: eigenvalues of $S$, where $l_i$: singular value
- $Y = (X - \bar{X}')V = UL$: PC scores

- the matrix $B$ that minimize

$$tr[(X - \bar{X} - B)(X - \bar{X} - B)']$$

over all $n \times p$ matrices $B$ having rank no greater than $k$ ($k < p$) is

$$B = UL_kV'$$

- $UL_k$ represents the first $k$ PC scores

- geometry of PC line (plane)

- the PC line (plane) minimizes the sum of squared orthogonal distances from each data point to the line (plane)

$${x_j} = (x'_j \hat{e}_1) \hat{e}_1 + (x'_j \hat{e}_2) \hat{e}_2 + \cdots + (x'_j \hat{e}_p) \hat{e}_p$$

$${y_j} = \hat{y}_{j1} \hat{e}_1 + \hat{y}_{j2} \hat{e}_2 + \cdots + \hat{y}_{jp} \hat{e}_p$$

- (c.f.) the least square line in regression minimizes the sum of vertical distances from the data point to the line
- drawbacks of PCA
  - PCA only utilizes information contained in the second moments
  - nonlinear structure may be missed
  - linear combination of variables may not be meaningful especially if the variables do not represent comparable quantities
  - outliers may distort the results

- Reading: Textbook, 8.1, 8.2, 8.3, 8.4, 8.5, Supplement 8A