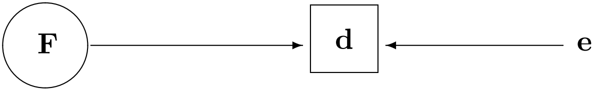
Principal Components[1]

STA431 Spring 2023

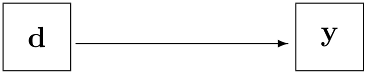
[1]See last slide for copyright information.

Principal Components Analysis is *not* Factor Analysis

- Factor analysis is the measurement model: $\mathbf{d} = \mathbf{\Lambda F} + \mathbf{e}$.



- Principal components are observable linear combinations: $\mathbf{y} = \mathbf{C}^\top \mathbf{d}$.



- Still, principal components and factor analysis have notable similarities and are frequently confused.

Data Reduction

- Suppose you have a large number of variables that are correlated with one another.
- Principal components analysis allows you to find a smaller set of linear combinations of the variables.
- There linear combinations may contain most of the variation in the original set.
- Use a few linear combinations in place of the entire data set.

Our Version

Standardized

- There are k observable variables, standardized: $z_j = \frac{x_j - \mu_j}{\sigma_j}$.
- $E(\mathbf{z}) = \mathbf{0}$, and $cov(\mathbf{z}) = \mathbf{\Sigma}$, a correlation matrix.
- $\mathbf{\Sigma} = \mathbf{C}\mathbf{D}\mathbf{C}^\top$

- $\mathbf{y} = \mathbf{C}^\top\mathbf{z}$ are the *principal components* of $\mathbf{z}$.
- A set of k linear combinations.

Rotation

- Because $\mathbf{C}\mathbf{C}^\top = \mathbf{I}$, $\mathbf{C}$ and $\mathbf{C}^\top$ are *orthogonal* matrices.
- Geometrically, multiplying a point by an orthogoanal matrix gives the location of the point in a new co-ordinate axis system, where the original axes have been *rotated*.
- For the multivariate normal, contours of constant probability density are ellipsoids.
- In principal components, the axes of the new co-ordinate system line up with the principal axes of the ellipsoids.

Mean and Covariance Matrix

Of principal components $\mathbf{y} = \mathbf{C}^\top \mathbf{z}$

$E(\mathbf{y}) = \mathbf{0}$, and

$$\begin{aligned} cov(\mathbf{y}) &= cov(\mathbf{C}^\top \mathbf{z}) \\ &= \mathbf{C}^\top cov(\mathbf{z}) \mathbf{C} \\ &= \mathbf{C}^\top \boldsymbol{\Sigma} \mathbf{C} \\ &= \mathbf{C}^\top \mathbf{C}\, \mathbf{D}\, \mathbf{C}^\top \mathbf{C} \\ &= \mathbf{D} \end{aligned}$$

So covariances of the principal components are all zero, and their variances are the eigenvalues.

$\mathbf{y} = \mathbf{C}^\top \mathbf{z} \iff \mathbf{z} = \mathbf{C}\mathbf{y}$

In scalar form,

$$\begin{aligned} z_1 &= c_{11}y_1 + c_{12}y_2 + \cdots + c_{1k}y_k \\ z_2 &= c_{21}y_1 + c_{22}y_2 + \cdots + c_{2k}y_k \\ \vdots & \qquad\qquad \vdots \\ z_k &= c_{k1}y_1 + c_{k2}y_2 + \cdots + c_{kk}y_k. \end{aligned}$$

So because the elements of $\mathbf{y}$ are uncorrelated,

$$\begin{aligned} Var(z_j) &= Var(c_{j1}y_1 + c_{j2}y_2 + \cdots + c_{jk}y_k) \\ &= c_{j1}^2 Var(y_1) + c_{j2}^2 Var(y_2) + \cdots + c_{jk}^2 Var(y_k) \\ &= c_{j1}^2 \lambda_1 + c_{j2}^2 \lambda_2 + \cdots + c_{jk}^2 \lambda_k = 1. \end{aligned}$$

Components of Variance

From

$$\begin{aligned} Var(z_j) &= Var(c_{j1}y_1 + c_{j2}y_2 + \cdots + c_{jk}y_k) \\ &= c_{j1}^2 Var(y_1) + c_{j2}^2 Var(y_2) + \cdots + c_{jk}^2 Var(y_k) \\ &= c_{j1}^2\lambda_1 + c_{j2}^2\lambda_2 + \cdots + c_{jk}^2\lambda_k = 1. \end{aligned}$$

we see

- The variance of z_j is decomposed into the part explained by y_1, the part explained by y_2, and so on.
- Specifically, y_1 explains $c_{j1}^2\lambda_1$ of the variance, y_2 explains $c_{j2}^2\lambda_2$ of the variance, etc..
- Because z_j is standardized, these are *proportions* of variance.

Squared Correlations

Using the fact that $cov(y_i, y_j) = 0$ for $i \neq j$,

$$\begin{aligned} Cov(z_i, y_j) &= Cov(c_{i1}y_1 + c_{i2}y_2 + \cdots + c_{ij}y_j + \cdots + c_{jk}y_k, y_j) \\ &= c_{ij}Cov(y_j, y_j) \\ &= c_{ij}\lambda_j. \end{aligned}$$

Then,

$$\begin{aligned} Corr(z_i, y_j) &= \frac{Cov(z_i, y_j)}{SD(z_i)SD(y_j)} \\ &= \frac{c_{ij}\lambda_j}{1\,\sqrt{\lambda_j}} = c_{ij}\sqrt{\lambda_j}, \end{aligned}$$

and the *squared* correlation between z_i and y_j is $c_{ij}^2\lambda_j$.

Squared correlation between z_i and y_j is $c_{ij}^2\lambda_j$

And using $Var(z_j) = c_{j1}^2\lambda_1 + c_{j2}^2\lambda_2 + \cdots + c_{jk}^2\lambda_k$

$$
\begin{array}{rcl}
Var(z_1) & = & c_{11}^2\lambda_1 + c_{12}^2\lambda_2 + \cdots + c_{1k}^2\lambda_k \\
Var(z_2) & = & c_{21}^2\lambda_1 + c_{22}^2\lambda_2 + \cdots + c_{2k}^2\lambda_k \\
\vdots & & \vdots \\
Var(z_k) & = & c_{k1}^2\lambda_1 + c_{k2}^2\lambda_2 + \cdots + c_{kk}^2\lambda_k.
\end{array}
$$

The pieces of variance being added up are the squared correlations between the original variables and the principal components.

A Matrix of Squared Correlations

Components of Variance

Element i, j is $Corr(z_i, y_j)^2$

	y_1	y_1	$\cdots$	y_k
z_1	$c_{11}^2\lambda_1$	$c_{12}^2\lambda_2$	$\cdots$	$c_{1k}^2\lambda_k$
z_2	$c_{21}^2\lambda_1$	$c_{22}^2\lambda_2$	$\cdots$	$c_{2k}^2\lambda_k$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
z_k	$c_{k1}^2\lambda_1$	$c_{k2}^2\lambda_2$	$\cdots$	$c_{kk}^2\lambda_k$

- If you add the entries in any row, you get one.
- Adding the entries in a column yields the total amount of variance in the original variables that is explained by that principal component.
- The sum of entries in column j is

$$\begin{aligned} \sum_{i=1}^{k} c_{ij}^2 \lambda_j &= \lambda_j \sum_{i=1}^{k} c_{ij}^2 \\ &= \lambda_j \cdot 1 = \lambda_j \end{aligned}$$

Meaning of the Eigenvalues of $\boldsymbol{\Sigma}$

	y_1	y_1	$\cdots$	y_k
z_1	$c_{11}^2\lambda_1$	$c_{12}^2\lambda_2$	$\cdots$	$c_{1k}^2\lambda_k$
z_2	$c_{21}^2\lambda_1$	$c_{22}^2\lambda_2$	$\cdots$	$c_{2k}^2\lambda_k$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
z_k	$c_{k1}^2\lambda_1$	$c_{k2}^2\lambda_2$	$\cdots$	$c_{kk}^2\lambda_k$
	λ_1	λ_2	$\cdots$	λ_k

The eigenvalues are both the variances of the principal components and the amounts of variance in the original variables that are explained by the respective principal components.

It gets better

A theorem says

- y_1 has the greatest possible variance of any linear combination whose squared weights add up to one.
- y_2 is the linear combination that has the greatest variance subject to the constraints that it's orthogonal to y_1 and its squared weights add to one.
- y_3 is the linear combination that has the greatest variance subject to the constraints that it's orthogonal to y_1 and y_2, and its squared weights add to one.
- And so on.
- It's a kind of optimality.

Data reduction

- If the correlations among the original variables are substantial, the first few eigenvalues will be relatively large.
- The data reduction idea is to retain only the first several principal components, the ones that contain most of the variation in the original variables.
- The expectation is that they will capture most of the *meaningful* variation.
- Conventional choice is to retain components with eigenvalues greater than one.

Sample Principal Components

- Of course we don't know $\mathbf{\Sigma}$, and we don't know means and standard deviations to standardize.
- So use the sample versions.

- $\mathbf{Z}$ is an $n \times k$ matrix of standardized variables.
- Independent (almost independent) random vectors are *row* vectors.
- Let $\mathbf{Y} = \mathbf{Z}\widehat{\mathbf{C}}$. Rows are sample principal components.
- All formulas apply to sample principal components, provided we use n in the denominators and not $n-1$.

- Principal components regression.

Copyright Information