FROM APPLICATION TO THEORY: SECTION II

PRINCIPAL COMPONENT ANALYSIS

REVIEW

Consider the linear combinations

- PCA projects *p*-dimensional data into a *q*-dimensional sub-space (*q*<=*p*)

$$
\begin{aligned}
Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \\
Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}
$$

$$\mathrm{var}(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{il}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_i$$

$$\mathrm{cov}(Y_i, Y_j) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{jl}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_j$$

$$\boxed{\mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}}$$

REVIEW

1^{st} Principal Component:

- The linear combination of X, i.e., Y_1 or PC_1, that has maximum variance, subject to the constrain that the sum of all e_{ij}^2 over $j=1,\ldots p$ is 1.

$$\mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}$$

More formally, select $\boldsymbol{e}_{11}, \boldsymbol{e}_{12}, \ldots, \boldsymbol{e}_{1p}$ to maximizes

$$\mathrm{var}(Y_1) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{1k}e_{1l}\sigma_{kl} = \mathbf{e}_1'\Sigma\mathbf{e}_1$$

Subject to:

$$\boxed{\mathbf{e}_1'\mathbf{e}_1 = \sum_{j=1}^{p} e_{1j}^2 = 1}$$

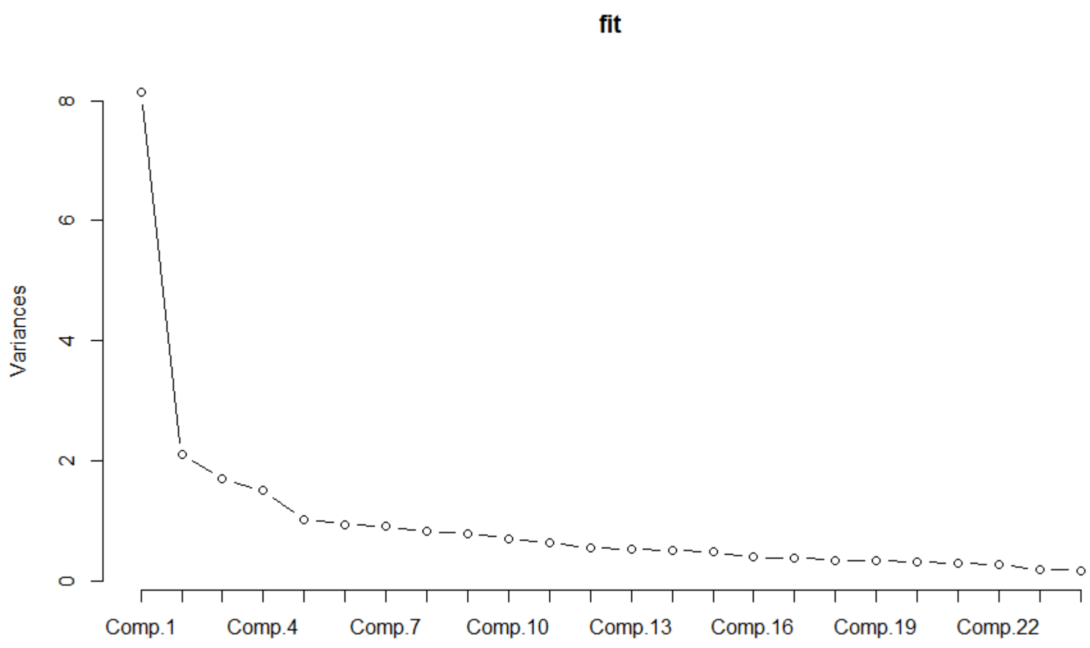
Correction: This is to ensure an unique answer

REVIEW

- First q Principal Component:
  - projected our p-dimensional data into a q-dimensional sub-space
  - We use the ratio of variance "explained" by the projected data to help us decide how many (q) PCs to retain (quantitatively and qualitatively)

REVIEW
HOW DO WE CHOOSE *Q*? - **VISUALIZATION**

- Screeplot – help to find the cutting point of choosing the number of PCs



REVIEW

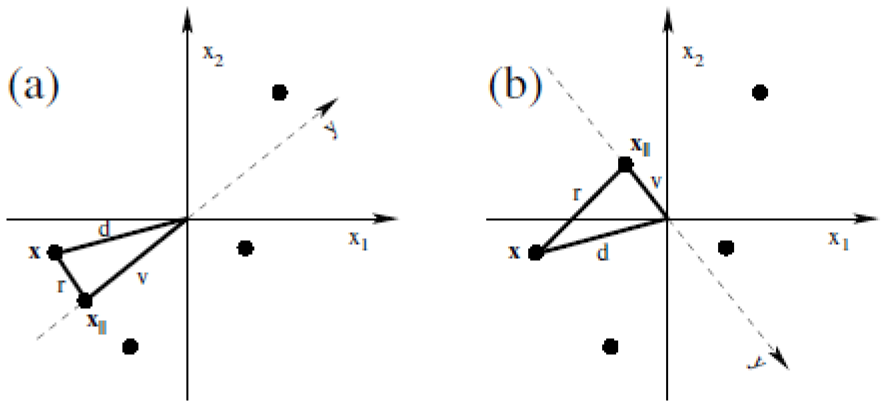
- First q Principal Component:
  - projected our p-dimensional data into a q-dimensional sub-space
  - We use the ratio of variance "explained" by the projected data to help us decide how many (q) PCs to retain (quantitatively and qualitatively)

Remember?

$$\boxed{R^2 = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{j=1}^{p} \lambda_j}}$$

λ: eigenvalues – we will talk about this today!

REVIEW: RECONSTRUCTION ERROR & VARIANCE



$$r^2 + v^2 = d^2$$

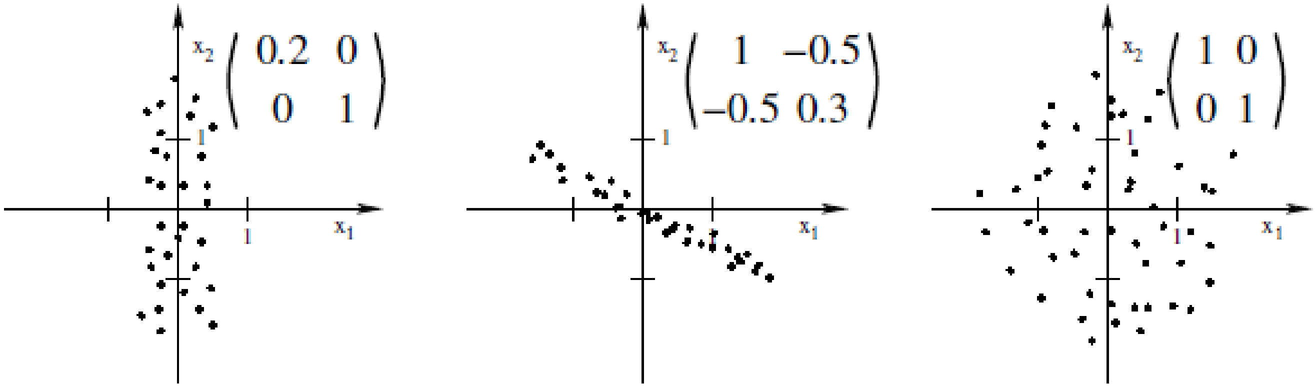
Reconstruction Error	Variance of the PC	Variance of the Data
Minimized	Maximized	Constant

REVIEW: DIRECTION OF MAXIMAL VARIANCE

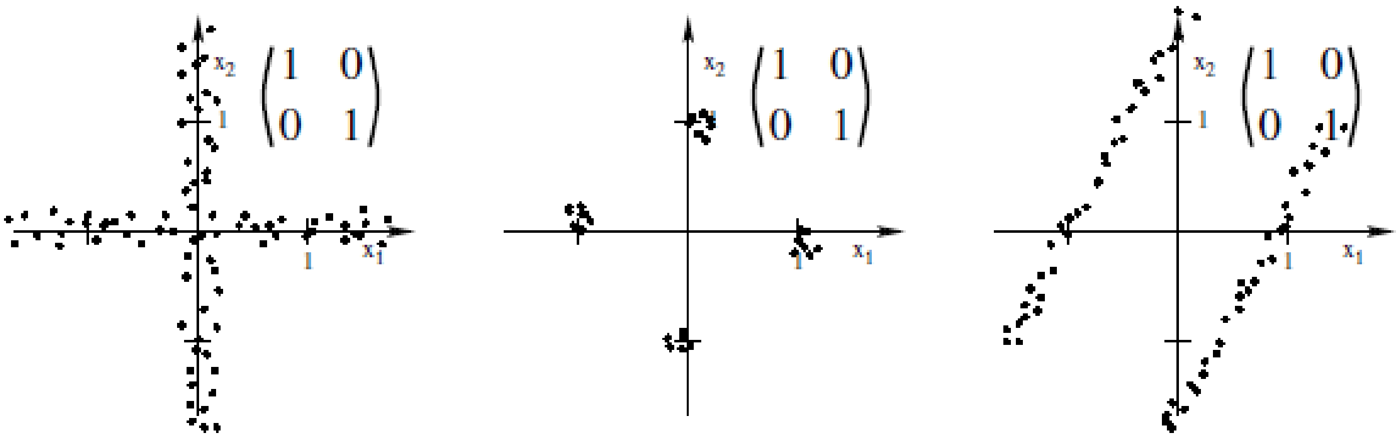
- **By Covariance Matrix**

III: $X = (x_1, x_2)^T \; assume \quad zero \quad mean$

$$C_{ij} = < x_i x_j >, \quad i = 1, 2; \; j = 1, 2$$



REVIEW: COVARIANCE ≠ DATA STRUCTURE



- ❑ The covariance matrix only gives you information about this general extent of the data, no higher-order structure of the data.

PCA DIMENSION REDUCTION

Input (high dimensional)

$x_1, x_2, \ldots. x_n$ points in R^p

Output (low dimensional)

$y_1, y_2, \ldots. y_n$ points in R^q $(q<<P)$

Projections

1. Assume inputs are centered: $\sum_i^n x_i = 0$ (x_i *is a vector*)
2. Given a <u>unit vector</u> u and a point x, the projection of x onto u is given by $X^T u$
3. Maximize projected variance:

$$\mathrm{var}(y) = \frac{1}{n}\sum_i (x_i^T u)^2 = \frac{1}{n}\sum_i u^T x_i x_i^T u$$

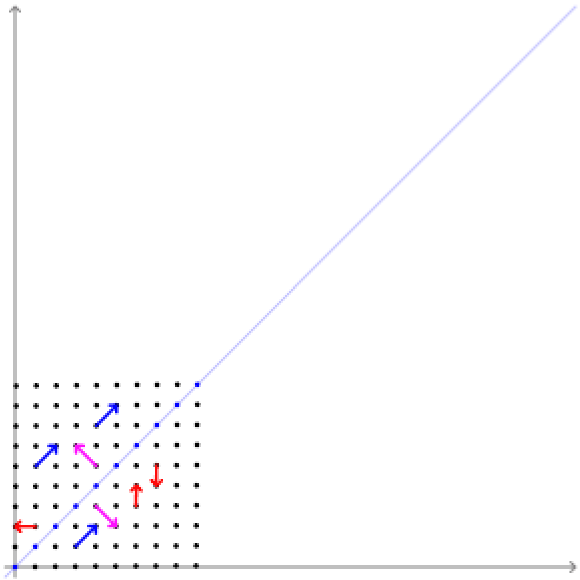
$$= u^T \left(\frac{1}{n}\sum_i x_i x_i^T\right) u = u^T C u$$

1. Assume inputs are centered: $\sum_i x_i = 0$
2. Given a unit vector u and a point x, the length of the projection of x onto u is given *by* $\boldsymbol{X}^T u$
3. Maximize projected variance: $$\text{var}(y) = \frac{1}{n}\sum_i (x_i^T u)^2 = \frac{1}{n}\sum_i u^T x_i x_i^T u$$ $$= u^T (\frac{1}{n}\sum_i x_i x_i^T) u = u^T C u$$
4. Minimize the sum of squared distances between all (x_i , y_i)

3 & 4 can be achieved simultaneously, we have a better explanation later.

5. If to a 1D subspace
   - Maximizing $u^T C u$ subject to $\|u\| = 1$, where we have $C = \frac{1}{n}\sum_i x_i x_i^T$ C is the empirical covariance matrix of the data, this gives the principal eigenvector of C

EIGENVECTORS



Graph source: wikipedia

How to find eigenvectors and eigenvalues?

6. If to a q-dimensional subspace
   - We need a collection of $u_1, \ldots u_q$ that are top q eigenvectors of C.
   - $u_1, \ldots u_q$ now form a new, orthogonal basis for the data.
   - We have a low-dimensional representation of **X**, by

$$y_i = \begin{bmatrix} u_1^T x_i \\ u_2^T x_i \\ \ldots \\ u_q^T x_i \end{bmatrix} \in \Re^q$$

TERMS TO INTERPRET PCA

- Eigenvectors:
  - The principal axes of maximum variance subspace
- Eigenvalues (λ)
  - The variance of projected inputs along principal axes
- Estimated dimensionality
  - The number of significant eigenvalues

$$R^2 = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

Note: λ here is a positive value different from when we derived the eigenvalues from the matrix on the blackbroad

PRACTICAL PART

PCA in R

R codes “WK3B_PCA in R Example.r”