Outline

Principal Component Analysis

Sometimes, we require $\|\boldsymbol{a}_1\| = 1$ and $\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle = 0$. Thus the problem is to find an interesting set of (orthogonal) direction vectors $\{\boldsymbol{a}_i : i = 1, \ldots, p\}$, where the projection scores of $\boldsymbol{X}$ onto $\boldsymbol{a}_i$ are useful.

Principal Component Analysis (PCA) is a linear dimension reduction technique that gives a set of direction vectors of maximal (projected) variances.

Take $d = 1$. PCA for the distribution of $\boldsymbol{X}$ finds $\boldsymbol{a}_1$ such that

$$\boldsymbol{a}_1 = \underset{\boldsymbol{a} \in \mathbb{R}^p, \|\boldsymbol{a}\|=1}{\operatorname{argmax}} \mathrm{Var}(Z_1(\boldsymbol{a})) \left(= \underset{\boldsymbol{a} \in \mathbb{R}^p, \|\boldsymbol{a}\|=1}{\operatorname{argmax}} \boldsymbol{a}' \mathrm{Var}(\boldsymbol{X}) \boldsymbol{a} \right),$$

where $Z_1(\boldsymbol{a}) = a_1 X_1 + \cdots + a_p X_p = \boldsymbol{a}' \boldsymbol{X}$.

The next section would be ......

1. PCA

2. Canonical Correlation Analysis

Linear dimension reduction

For a random vector $\boldsymbol{X} \in \mathbb{R}^p$, consider reducing the dimension from p to d, i.e., p variables $(X_1, \ldots, X_p)^T$ to a set of *most interesting* d variables. Here, $1 \leq d \leq p$.

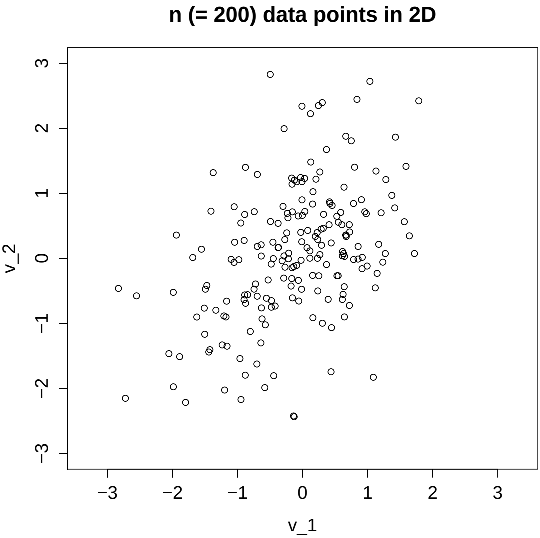
- Best subset?
- Linear dimension reduction: Construct d variables $Z_1, \ldots, Z_d$ as linear combinations of $X_1, \ldots, X_p$, i.e.

$$Z_i = a_{i1}X_1 + \cdots + a_{ip}X_p = \boldsymbol{a}_i'\boldsymbol{X} \quad (i = 1, \ldots, d),$$

with $\boldsymbol{a}_i \in \mathbb{R}^p$.
Linear dimension reduction seeks a sequence of such Z_i, or equivalently a sequence of $\boldsymbol{a}_i$, where the random variables (Z_is) are most important among all other choices.

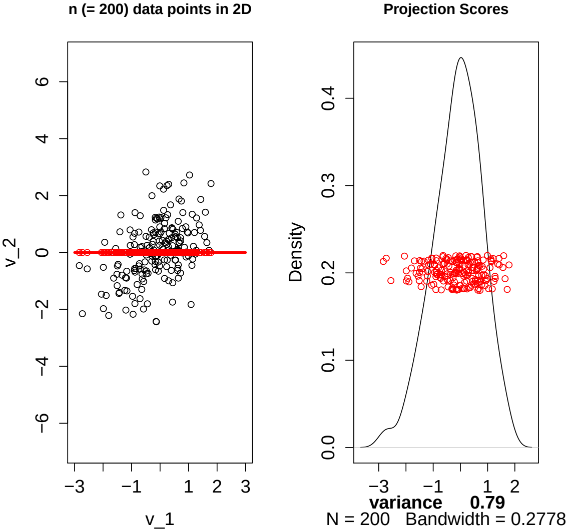
Geometric understanding of PCA for point cloud

PCA is best understood with a point cloud. Take a look at 2D example.



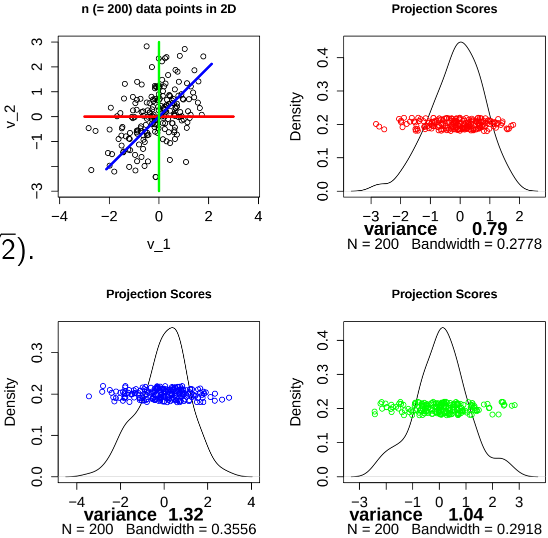
Geometric understanding of PCA for point cloud

Take $\boldsymbol{a} = (1, 0)'$.

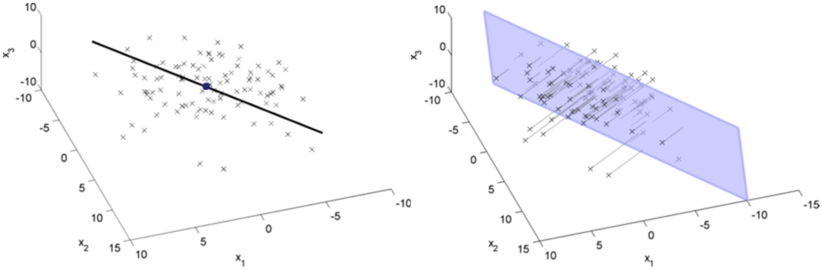


Geometric understanding of PCA for point

Take $\boldsymbol{a} = (1,0)'$, $(0,1)', (1/\sqrt{2}, 1/\sqrt{2})$. Which one is better?



Geometric understanding of PCA for 3D point cloud



A 3D point cloud. Mean. 1st PC direction (maximizing variance of projections) explains the cloud best. 1st and 2nd directions form a plane.

Formulation of population PCA–1

Suppose a random vector $\boldsymbol{X}$ with mean μ, covariance Σ (not necessarily normal).

The first principal component (PC) direction vector is the unit vector $\boldsymbol{u}_1 \in \mathbb{R}^p$ that maximizes the variance of $\boldsymbol{u}_1'\boldsymbol{X}$ when compared to other unit vectors, i.e.,

$$\boldsymbol{u}_1 = \underset{\boldsymbol{u} \in \mathbb{R}^p, \|\boldsymbol{u}\|=1}{\operatorname{argmax}} \mathrm{Var}(\boldsymbol{u}'\boldsymbol{X}).$$

- $\boldsymbol{u}_1 = (u_{11}, \ldots, u_{1p})'$ is the first PC direction vector, sometimes called *loading vector*.
- $(u_{11}, \ldots, u_{1p})$ are loadings of the 1st PC.
- $Z_1 = u_{11}X_1 + \cdots + u_{1p}X_p = \boldsymbol{u}_1'\boldsymbol{X}$ is the first PC score or the first principal component (random variable).
- $\lambda_1 = \mathrm{Var}(\boldsymbol{u}'\boldsymbol{X}) = \mathrm{Var}(Z_1)$ is the variance explained by the first PC.

Formulation of population PCA–2

The second PC direction is the unit vector $\boldsymbol{u}_2 \in \mathbb{R}^p$ that

- maximizes the variance of $\boldsymbol{u}_2'\boldsymbol{X}$;
- is orthogonal to the first PC direction $\boldsymbol{u}_1$.

That is,

$$\boldsymbol{u}_2 = \underset{\boldsymbol{u}\in\mathbb{R}^p, \|\boldsymbol{u}\|=1, \boldsymbol{u}'\boldsymbol{u}_1=0}{\operatorname{argmax}} \mathrm{Var}(\boldsymbol{u}'\boldsymbol{X}).$$

- $\boldsymbol{u}_2 = (u_{21}, \ldots, u_{2p})'$ is the second PC direction vector, and is the vector of the 2nd set of loadings.
- $Z_2 = \boldsymbol{u}_2'\boldsymbol{X}$ is the second principal component.
- $\lambda_2 = \mathrm{Var}(Z_2)$ is the variance explained by the second PC, and $\lambda_1 \geq \lambda_2$.
- $\mathrm{Corr}(Z_1, Z_2) = 0$.

Formulation of population PCA–(3,4,...p)

Given the first $k-1$ PC directions $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$, the kth PC direction is the unit vector $\boldsymbol{u}_k \in \mathbb{R}^p$ that

- maximizes the variance of $\boldsymbol{u}_k' \boldsymbol{X}$;
- is orthogonal to the 1st to $(k-1)$th PC directions $\boldsymbol{u}_j$.

That is,

$$\boldsymbol{u}_k = \underset{\substack{\boldsymbol{u} \in \mathbb{R}^p, \|\boldsymbol{u}\|=1 \\ \boldsymbol{u}'\boldsymbol{u}_j = 0, j=1,\ldots,k-1}}{\operatorname{argmax}} \mathrm{Var}(\boldsymbol{u}'\boldsymbol{X}).$$

- $\boldsymbol{u}_k = (u_{k1}, \ldots, u_{kp})'$ is the kth PC direction vector, and is the vector of the kth loadings.
- $Z_k = \boldsymbol{u}_k' \boldsymbol{X}$ is the kth principal component.
- $\lambda_k = \mathrm{Var}(Z_k)$ is the variance explained by the k PC score, and $\lambda_1 \geq \cdots \geq \lambda_{k-1} \geq \lambda_k$.
- $\mathrm{Corr}(Z_i, Z_j) = 0$ for all $i \neq j \leq k$.

Relation to eigen-decomposition of Σ

Recall the eigen-decomposition of the symmetric positive definite $\Sigma = \boldsymbol{U}\Lambda\boldsymbol{U}'$ with

- $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p]$ orthogonal,
- $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \ldots \geq \lambda_p$,
- $\Sigma\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i$.

In next two slides we show that:

1. The kth eigenvector $\boldsymbol{u}_k$ is the kth PC direction vector.
2. The kth eigenvalue λ_k is the variance explained by the kth principal component.
3. PC directions are both orthogonal $\boldsymbol{u}_i'\boldsymbol{u}_j = 0$ $(i \neq j)$ and Σ-orthogonal

$$\boldsymbol{u}_i'\Sigma\boldsymbol{u}_j = 0 \iff \mathrm{Cov}(Z_i, Z_j) = 0 \quad (i \neq j).$$

Relation to eigen-decomposition of Σ

The first PC direction problem is to maximize $\text{Var}(\boldsymbol{u}'\boldsymbol{X})$ with the constraint $\boldsymbol{u}'\boldsymbol{u} = 1$. Using Lagrange multiplier λ, the problem of maximization is the same as finding a stationary point of

$$\begin{aligned}\Phi(\boldsymbol{u}, \lambda) &= \text{Var}(\boldsymbol{u}'\boldsymbol{X}) - \lambda(\boldsymbol{u}'\boldsymbol{u} - 1) \\ &= \boldsymbol{u}'\Sigma\boldsymbol{u} - \lambda(\boldsymbol{u}'\boldsymbol{u} - 1).\end{aligned}$$

The stationary point solves the following:

$$\frac{1}{2}\frac{\partial}{\partial \boldsymbol{u}}\Phi(\boldsymbol{u}, \lambda) = \boldsymbol{\Sigma}\boldsymbol{u} - \lambda\boldsymbol{u} = \boldsymbol{0},$$

which leads to

$$\lambda = \boldsymbol{u}'\Sigma\boldsymbol{u} = \text{Var}(\boldsymbol{u}'\boldsymbol{X}), \quad \Sigma\boldsymbol{u} = \lambda\boldsymbol{u}. \tag{1}$$

Any eigenvector-eigenvalue pair $(\boldsymbol{u}_i, \lambda_i), (i = 1, \ldots, p)$ satisfies the second eq. in (1). It is clear that the first PC direction is the first eigenvector $\boldsymbol{u}_1$, as it gives the largest variance $\lambda_1 = \boldsymbol{u}_1'\Sigma\boldsymbol{u}_1 = \text{Var}(\boldsymbol{u}_1'\boldsymbol{X}) \geq \lambda_j \ (j > 1)$.

Relation to eigen-decomposition of Σ

For the kth PC direction, we form a Lagrangian function

$$\Phi(\boldsymbol{u}, \lambda, \gamma_1^k) = \boldsymbol{u}'\Sigma\boldsymbol{u} - \lambda(\boldsymbol{u}'\boldsymbol{u} - 1) - \textstyle\sum_{j=1}^{k-1} 2\gamma_j \boldsymbol{u}_j'\boldsymbol{u},$$

given the first $k-1$ PC directions. The derivative of Φ, equated to zero, is then

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial \boldsymbol{u}}\Phi(\boldsymbol{u}, \lambda, \gamma_1^k) &= \boldsymbol{\Sigma u} - \lambda\boldsymbol{u} - \textstyle\sum_{j=1}^{k-1} \gamma_j \boldsymbol{u}_j = \boldsymbol{0}, \\ \frac{\partial}{\partial \gamma_j}\Phi(\boldsymbol{u}, \lambda, \gamma_1^k) &= \boldsymbol{u}_j'\boldsymbol{u} = 0. \end{aligned} \tag{2}$$

We have $\gamma_j = \boldsymbol{u}_j'\Sigma\boldsymbol{u} = 0$ (since $\Sigma\boldsymbol{u}_j = \lambda_j\boldsymbol{u}_j$), thus

$$\lambda = \boldsymbol{u}'\Sigma\boldsymbol{u} = \mathrm{Var}(\boldsymbol{u}'\boldsymbol{X}), \quad \Sigma\boldsymbol{u} = \lambda\boldsymbol{u}. \tag{3}$$

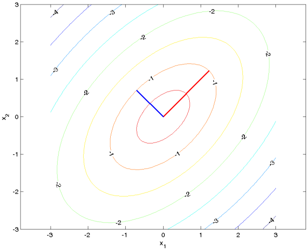
The kth to the last eigen-pairs $(\boldsymbol{u}_i, \lambda_i), (i = k, \ldots, p)$ all satisfy both (3) and (2). Thus, the kth PC direction is $\boldsymbol{u}_k$, as it gives the largest variance $\lambda_k = \boldsymbol{u}_k'\Sigma\boldsymbol{u}_k$.

Example: Principal components of bivariate normal distribution

Consider $N_2(\mathbf{0}, \Sigma)$ with

$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{bmatrix}',$$

The PC directions $\boldsymbol{u}_i$ are the principal axes of ellipsoids, representing the density of MVN, with lengths given by $\sqrt{\lambda_i}$.



Sample PCA

Given multivariate data $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]_{p \times n}$, the *sample PCA* sequentially finds orthogonal directions of maximal (projected) sample variance.

1. For $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$, the centered data matrix is

$$\tilde{\mathbf{X}} = [\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n] = \mathbf{X}(\mathbb{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}').$$

2. The sample variance matrix is then

$$\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'.$$

3. Eigen-decomposition of $\mathbf{S} = \widehat{\boldsymbol{U}}\widehat{\Lambda}\widehat{\boldsymbol{U}}'$ leads to the kth sample PC direction ($\hat{\boldsymbol{u}}_k$) and the variance of the kth sample score ($\hat{\lambda}_k$).

Sample PCA

It can be checked that the eigenvector $\hat{\boldsymbol{u}}_k$ of $\mathbf{S}$ satisfies

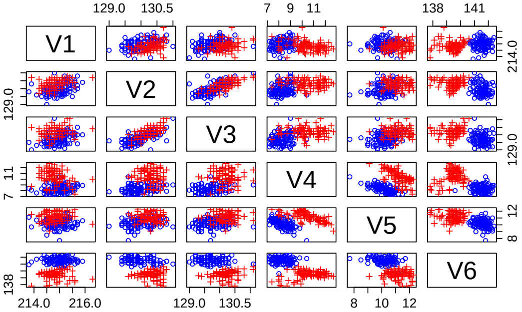
$$\hat{\boldsymbol{u}}_k = \underset{\substack{\boldsymbol{u}\in\mathbb{R}^p, \|\boldsymbol{u}\|=1 \\ \boldsymbol{u}'\hat{\boldsymbol{u}}_j=0, j=1,\ldots,k-1}}{\operatorname{argmax}} \boldsymbol{u}'\mathbf{S}\boldsymbol{u}$$

- $\hat{\boldsymbol{u}}_k = (\hat{u}_{k1}, \ldots, \hat{u}_{kp})'$ is the kth sample PC direction vector, and is the vector of the kth loadings.
- $\boldsymbol{z}_{(k)} = (z_{(k)i}, \ldots, z_{(k)n})_{1\times n} = \hat{\boldsymbol{u}}_k'\tilde{\boldsymbol{X}}$ is the of kth score vector, where $z_{(k)i} = \hat{\boldsymbol{u}}_k'\tilde{\boldsymbol{x}}_i$.
- $\lambda_k =$ the sample variance of $\{z_{(k)i} : i = 1, \ldots, n\}$.
- $\boldsymbol{u}'\mathbf{S}\boldsymbol{u} = (n-1)^{-1}(\boldsymbol{u}'\tilde{\mathbf{X}})(\tilde{\mathbf{X}}'\boldsymbol{u}) = (n-1)^{-1}\sum_{i=1}^{n}(\boldsymbol{u}'\tilde{\boldsymbol{x}}_i)^2$, which is the sample variance of $\{\boldsymbol{u}'\boldsymbol{x}_i, i = 1, \ldots, n\}$

The information of the first two principal components is all contained in the score vectors $(\boldsymbol{z}_{(1)}, \boldsymbol{z}_{(2)})$, the 2D scatters of which is thus most informative with the largest total variance.

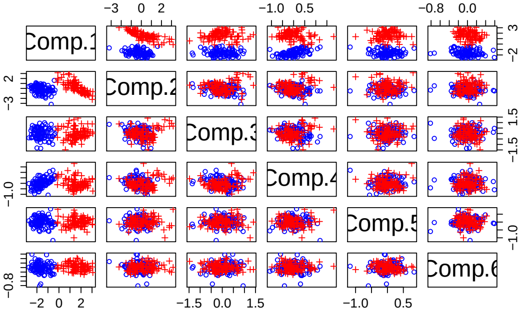
Example: Swiss Bank Note Data

Swiss Bank Note data from HS. $p = 6, n = 200$. Genuine (blue) and fake (red) samples in original measurements.



Example: Swiss Bank Note Data

Scatters of principal components



How many components to keep? (1)

The criterion for PCA is a high variance in the principal components. The question involves "how much the PCs explain the variation within the whole data."

1. Total variance in $\mathbf{X}$ is the sum of all marginal variances

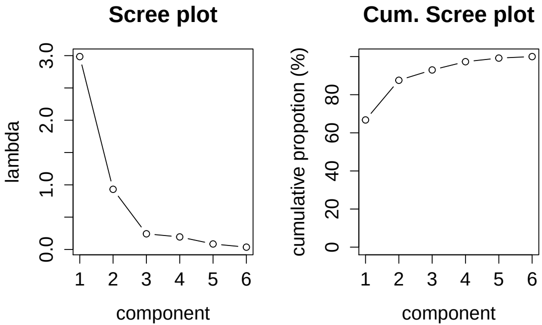
$$\begin{aligned}\sum_{k=1}^{p} \mathrm{Var}(\{x_{ki} : i = 1, \ldots, n\}) &= \mathrm{Trace}(\Sigma) \\ &= \sum_{k=1}^{p} \mathrm{Var}(\{z_{(k)i} : i = 1, \ldots, n\}) = \sum_{k=1}^{p} \lambda_k.\end{aligned}$$

2. Variance of the kth scores: $\mathrm{Var}(\{z_{(k)i} : i = 1, \ldots, n\}) = \lambda_k$.
3. Total variance in the 1st–kth PCs: $\lambda_1 + \ldots + \lambda_k$.

For heuristics made from these quantities, see next slide:

Example: Swiss Bank Note Data–scree plot

In scree plot (k, λ_k), we look for an elbow.
In cumulative scree plot, (proportion of variance explained, $(k, \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j})$), use 90% as a cutoff.



How many components to keep? (2)

1. Kaiser's rule (of thumb): Retain PCs 1–k satisfying

$$\lambda_k > \bar{\lambda} = \frac{1}{p} \sum_{j=1}^{p} \lambda_j.$$

   Tends to choose fewer components.

2. Likelihood ratio testing on null hypothesis

$$H_0(k) : \lambda_{k+1} = \cdots = \lambda_p,$$

   where k components are used if $H_0(k)$ is not rejected at a specified level.

Which variables are most responsible for the principal components?

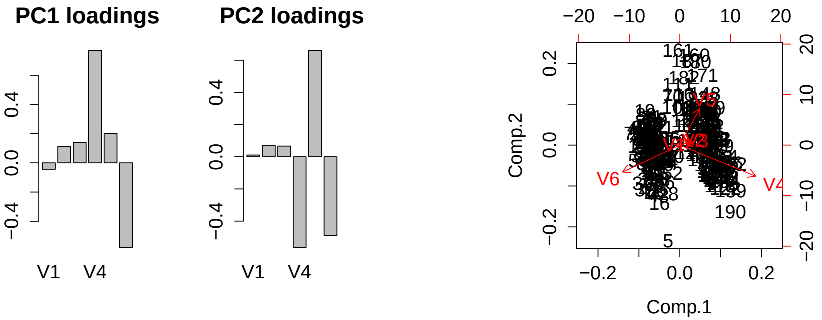
- Loadings of principal component direction.
- Biplot - scatterplot of PC1 and PC2 scores, overlaid with p vectors each representing the loadings of the first two PC directions.

In the Swiss Bank Note Data, the loadings are

```
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
V1               -0.326  0.562  0.753       
V2  0.112        -0.259  0.455 -0.347 -0.767
V3  0.139        -0.345  0.415 -0.535  0.632
V4  0.768 -0.563 -0.218 -0.186              
V5  0.202  0.659 -0.557 -0.451  0.102       
V6 -0.579 -0.489 -0.592 -0.258              
```

Example: Swiss Bank Note Data–biplot

Which measurements are most responsible for the first two principal components?



Computation of PCA

PCA is either computed using *eigenvalue decomposition* of $\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ or using the *singular value decomposition* of $\tilde{\boldsymbol{X}}$.

Eigen-decomposition of $\mathbf{S}$

For $\mathbf{S} = \boldsymbol{U}\Lambda\boldsymbol{U}'$,

1. PC directions $\boldsymbol{u}_k$ (eigenvectors)
2. Variance of PC (scores) λ_k (eigenvalues)
3. Matrix of principal component scores

$$\boldsymbol{V} = \boldsymbol{U}'\mathbf{X} = \begin{bmatrix} \boldsymbol{z}_{(1)} \\ \vdots \\ \boldsymbol{z}_{(p)} \end{bmatrix}.$$

Computation of PCA

Singular value decomposition (SVD) of $\tilde{\mathbf{X}}$

The singular value decomposition (SVD) of $p \times n$ matrix $\tilde{\mathbf{X}}$ has the form

$$\tilde{\mathbf{X}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'.$$

- The left singular vectors $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p]_{p \times p}$ and the right singular vectors $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p]_{n \times p}$ are orthogonal ($\boldsymbol{U}'\boldsymbol{U} = \mathbb{I}_p, \boldsymbol{V}'\boldsymbol{V} = \mathbb{I}_p$).
- The columns of $\boldsymbol{U}$ span the column space of $\tilde{\mathbf{X}}$; the columns of $\boldsymbol{V}$ span the row space.
- $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_p)$, $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ are the singular values of $\tilde{\mathbf{X}}$.

Computation of PCA

SVD of $\tilde{\mathbf{X}}$

The SVD of $\tilde{\mathbf{X}} = \boldsymbol{UDV}'$. Then

$$\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}' = \frac{1}{n-1}\boldsymbol{UDV}'\boldsymbol{VDU}' = \boldsymbol{U}\mathrm{diag}(\frac{1}{n-1}d_j^2)\boldsymbol{U}'$$

1. PC directions $\boldsymbol{u}_k$ (left singular vectors)
2. Variance of PC (scores) $\frac{1}{n-1}d_j^2$ (singular values^2)
3. Matrix of principal component scores (right singular vectors)

$$\begin{bmatrix} \boldsymbol{z}_{(1)} \\ \vdots \\ \boldsymbol{z}_{(p)} \end{bmatrix} = \boldsymbol{Z} = \boldsymbol{U}'\tilde{\mathbf{X}} = \boldsymbol{DV}' = \begin{bmatrix} d_1\boldsymbol{v}_1' \\ \vdots \\ d_p\boldsymbol{v}_p' \end{bmatrix}$$

NOTE: we are working with the centered $\tilde{\mathbf{X}}$ here, not $\mathbf{X}$!!

PCA in R

The standard data format is the $n \times p$ data frame or matrix `x`.
To perform PCA by eigen decomposition:

```
spr <-princomp(x)
U<-spr$loadings
L<-(spr$sdev)^2
Z <-spr$scores
```

To perform PCA by singular value decompositoin

```
gpr <- prcomp(x)
U <- gpr$rotation
L <- (gpr$sdev)^2
Z <- gpr$x
```

Correlation PCA

- PCA is not scale invariant.
- Correlation matrix of a random vector $\mathbf{X}$ is given by

$$\boldsymbol{R} = D_{\Sigma}^{-\frac{1}{2}} \Sigma D_{\Sigma}^{-\frac{1}{2}},$$

  where D_{Σ} is the $p \times p$ diagonal matrix consisting of diagonal elements of Σ.
- Correlation PCA: The PC directions and variance of PC scores are obtained by eigen-decomposition of $\boldsymbol{R} = \boldsymbol{U}_R \Lambda_R \boldsymbol{U}'_R$.
- Preferred if measurements are not commensurate (e.g. $X_1 =$ household income, $X_2 =$ years in school).

PCA for Olivetti Faces data

Olivetti Faces data

- Obtained from `http://www.cs.nyu.edu/~roweis/data.html`.
- Grayscale faces 8 bit [0-255], a few (10) images of several (40) different people.
- 400 total images, 64x64 size.
- From the Oivetti database at ATT.

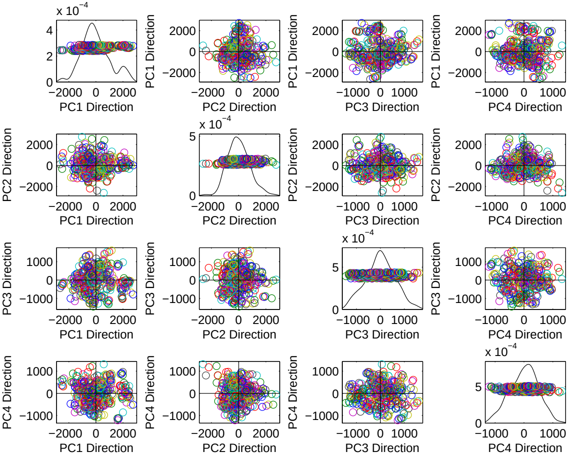
4 samples, 64 x 64 pixels, each pixel has a value 0−255 (darker−lighter)

An image is a matrix-valued datum. In Olivetti Faces data, the matrix is of size 64×64, with each pixel having values between [0-255]. The matrix, corresponding one observation, is vecterized (vec'd) by stacking each column into one long vector of size $d = 4096 = 64 \times 64$.

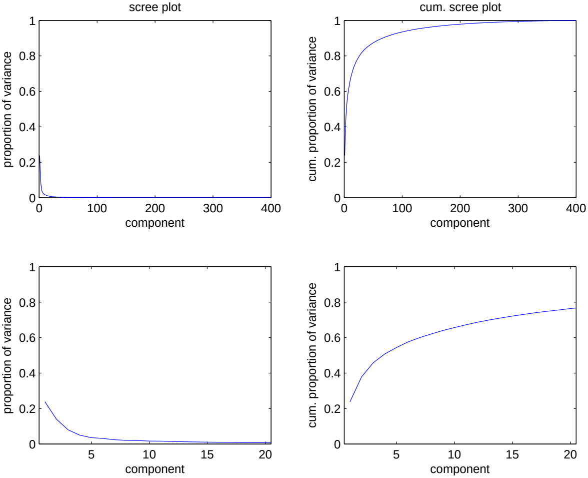
So, $\boldsymbol{x}_1$ is a $d \times 1$ vector corresponding to

PCA is applied to the data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$.

Olivetti Faces data–Major components

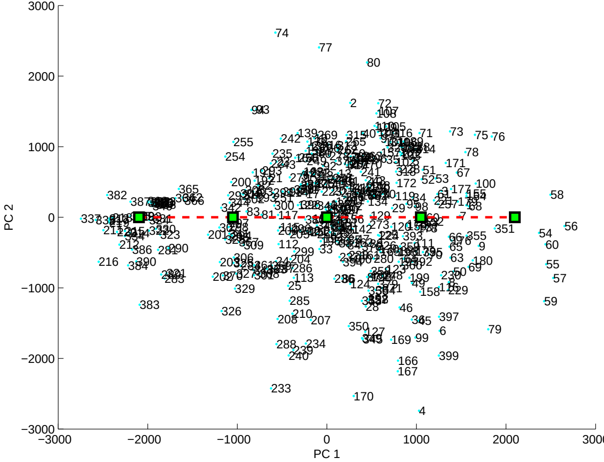


Olivetti Faces data–Scree plots

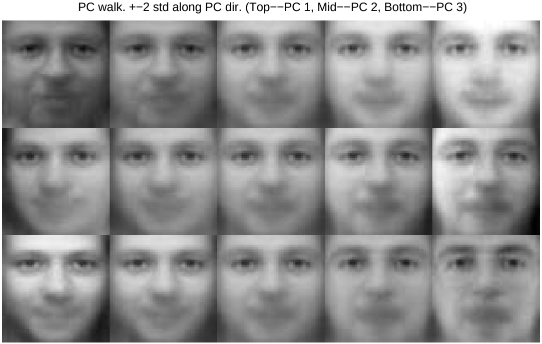


Olivetti Faces data–Interpretation

Examine the mode of variation by walking along the PC direction from the mean. Here, $\pm 1, 2$ standard deviations of $Z_{(k)}$.



Olivetti Faces data–Interpretation (Eigenfaces)



PC1 $\sim$ darker to lighter face
PC2 $\sim$ feminin to masculine face
PC3 $\sim$ oval to rectangle face

Olivetti Faces data–Interpretation (Eigenfaces–Computation)

Vectorized data in $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, with mean $\bar{\boldsymbol{x}}$. Denote $(\boldsymbol{u}_j, \lambda_j)$: the jth PC direction and PC variance.
Walk along PC j direction and examine at position $s = \pm 2, \pm 1, 0$ by

1. Reconstruction at s: $\boldsymbol{w}_s = \bar{\boldsymbol{x}} \pm s\sqrt{\lambda_j}\boldsymbol{u}_j$
2. Convert to image by reshaping the 4096×1 vector $\boldsymbol{w}_s$ into 64×64 matrix $\boldsymbol{W}_s$.

Olivetti Faces data–Interpretation (Eigenfaces–Computation)

Vectorized data in $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, with mean $\bar{\boldsymbol{x}}$. Denote $(\boldsymbol{u}_j, \lambda_j)$: the jth PC direction and PC variance.
Walk along PC j direction and examine at position $s = \pm 2, \pm 1, 0$ by

1. Reconstruction at s: $\boldsymbol{w}_s = \bar{\boldsymbol{x}} \pm s\sqrt{\lambda_j}\boldsymbol{u}_j$
2. Convert to image by reshaping the 4096×1 vector $\boldsymbol{w}_s$ into 64×64 matrix $\boldsymbol{W}_s$.

Olivetti Faces data–Reconstruction of original data

Approximation to the original data matrix:

$$\boldsymbol{x}_i = \bar{\boldsymbol{x}} + \sum_{j=1}^{p} z_{(j)i}\boldsymbol{u}_j, \quad (i = 1, \ldots, n)$$

Approximation of the original observation $\boldsymbol{x}_i$ by the first $m < p$ principal components:

$$\hat{\boldsymbol{x}}_i = \bar{\boldsymbol{x}} + \sum_{j=1}^{m} z_{(j)i}\boldsymbol{u}_j,$$

- The larger m, the better approximation by $\hat{\boldsymbol{x}}_i$.
- The smaller m, the more succinct dimension reduction of $\boldsymbol{X}$.

See some mathematical explanations in the next page.

Olivetti Faces data–Reconstruction of original data

Recall

1. $\tilde{\mathbf{X}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$
2. $\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{V}'$
3. $\tilde{\mathbf{X}} = \boldsymbol{U}\boldsymbol{Z}$
4. $\tilde{\boldsymbol{x}}_i = \boldsymbol{U}\boldsymbol{z}_i = \sum_{j=1}^{p} \boldsymbol{u}_j z_{(j)i}$
   In a coordinate system with $\{\boldsymbol{u}_i, i = 1, \ldots, n\}$ as the p basis vectors, $z_{(j)i}$ is the jth coordinate for the ith observation
   $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$.

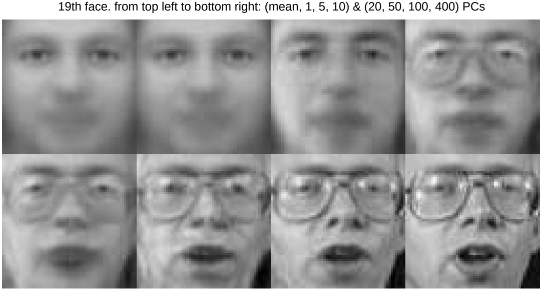
Reconstruction of original face

Observation index $i = 5$.

5th face. from top left to bottom right: (mean, 1, 5, 10) & (20, 50, 100, 400) PCs

Reconstruction of original face

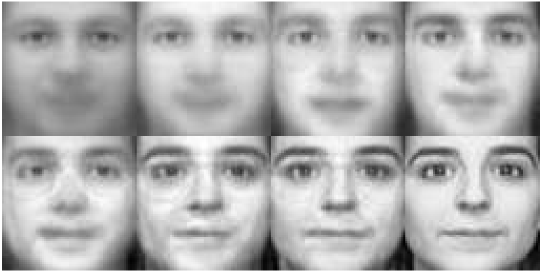
Observation index $i = 19$.



Reconstruction of original face

Observation index $i = 100$.

100th face. from top left to bottom right: (mean, 1, 5, 10) & (20, 50, 100, 400) PCs



Olivetti Faces data
Reconstruction of original face

- Human eyes require > 50 principal components to see resemblance between $\hat{\boldsymbol{x}}_i$ and $\boldsymbol{x}_i$.
- Corresponds to about 90 percent of variance explained in PCs.
- Subjective and heuristic decision on "how many components to use"
- Reconstruction by PCA most useful when
  - each datum is visually represented (rather than being just numbers).
  - for example: data objects are images, functions, shapes.

PCA as a mean of dimension reduction

- We can use only the first d PCs to approximately represent the data. Instead of $\mathbf{X}_{p \times n}$, we store the data as $\mathbf{Z}_{d \times n}$.
- This is essentially a dimension reduction approach. However,
  1. Unsupervised learning (no information on Y).
  2. Each PC (new variable) is a linear combination of p variables. Hard to interpret.
  3. Eigen-decomposition/SVD are problematic when $p \gg n$.

Some open questions

- PCA consistency
  1. p fixed $n \to \infty$
  2. n fixed $p \to \infty$
  3. n and $p \to \infty$
- Sparse PCA
  1. Many zeros in $\boldsymbol{u}_k$

The next section would be ......

Dimension reduction for two sets of random variables

When distinction between explanatory and response variables are not so clear, an analysis dealing with the two sets of variables in a symmetric manner is desired.

Examples

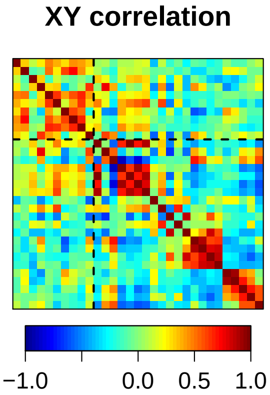
1. Relationship between genes expressions and biological variables;
2. Relation between two sets of psychological tests, each with multidimensional measurements.

Example: `nutrimouse` dataset

- a nutrition study in the mouse
- Reference: `http://www6.toulouse.inra.fr/toxalim/pharmaco-moleculaire/acceuil.html`
  Pascal Martin from the Toxicology and Pharmacology Laboratory (French National Institute for Agronomic Research).
- Obtained from R package, CCA.
- $n = 40$ mice, each with $r = 120$ gene expression levels, associated with nutrition problem, and $s = 21$ measurements of concentrations of 21 hepatic fatty acids.

$$\boldsymbol{X}_{120\times 40}, \quad \boldsymbol{Y}_{21\times 40}$$

Example: `nutrimouse` dataset–Objective

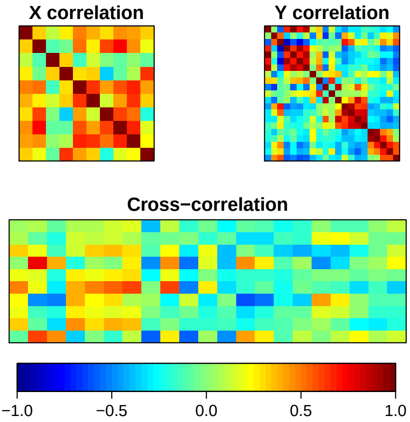


Focus on the first 10 gene expression levels (Dimension r of the first set of variables is now 10).

$$\boldsymbol{X}_{10\times 40}, \quad \boldsymbol{Y}_{21\times 40}$$

Example: `nutrimouse` dataset–Objective

- Interested in dimension reduction of $\boldsymbol{X}$ and $\boldsymbol{Y}$, while keeping the important association between $\boldsymbol{X}$ and $\boldsymbol{Y}$.
- Find linear dimension reduction of $\boldsymbol{X}$ and $\boldsymbol{Y}$ using the cross-correlation matrix



Canonical Correlation Analysis (CCA)

CCA seeks to identify and quantify the associations between two sets of variables.
Given two random vectors $\boldsymbol{X} \in \mathbb{R}^r$ and $\boldsymbol{Y} \in \mathbb{R}^s$ ($r = s$ or $r \neq s$), consider *linear dimension reduction* of each of two random vectors,

$$\xi = \boldsymbol{g}'\boldsymbol{X} = g_1 X_1 + \cdots + g_r X_r,$$
$$\omega = \boldsymbol{h}'\boldsymbol{Y} = h_1 Y_1 + \cdots + h_s Y_s.$$

CCA finds the random variables (ξ, ω) or the projection vectors $(\boldsymbol{g}, \boldsymbol{h})$ that give *maximal correlation* between ξ and ω,

$$\text{Corr}(\xi, \omega) = \frac{\text{Cov}(\xi, \omega)}{\sqrt{\text{Var}(\xi)\text{Var}(\omega)}}$$

Population CCA – 1

Denote

$$\Sigma_{11} = \text{Cov}(\boldsymbol{X}), \Sigma_{22} = \text{Cov}(\boldsymbol{Y})$$
$$\Sigma_{12} = \text{Cov}(\boldsymbol{X}, \boldsymbol{Y}), \Sigma_{21} = \text{Cov}(\boldsymbol{Y}, \boldsymbol{X}).$$

The first set of projection vectors

$$(\boldsymbol{g}_1, \boldsymbol{h}_1) = \underset{\boldsymbol{g}\in\mathbb{R}^r, \boldsymbol{h}\in\mathbb{R}^s}{\text{argmax}} \text{Corr}(\boldsymbol{g}'\boldsymbol{X}, \boldsymbol{h}'\boldsymbol{Y}) \tag{4}$$
$$\underset{\boldsymbol{g}\in\mathbb{R}^r, \boldsymbol{h}\in\mathbb{R}^s}{\text{argmax}} \frac{\boldsymbol{g}'\Sigma_{12}\boldsymbol{h}}{\sqrt{\boldsymbol{g}'\Sigma_{11}\boldsymbol{g}\,\boldsymbol{h}'\Sigma_{22}\boldsymbol{h}}}. \tag{5}$$

- $(\boldsymbol{g}_1, \boldsymbol{h}_1)$ Canonical correlation vectors.
- $(\xi_1 = \boldsymbol{g}_1'\boldsymbol{X}, \omega_1 = \boldsymbol{h}_1'\boldsymbol{Y})$ are called canonical variates (like PC).
- $\rho_1 = \text{Corr}(\xi_1, \omega_1)$ is the largest canonical correlation.

Population CCA – 2,3,...

The kth set of projection vectors, given $(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{k-1})$ and $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{k-1})$, are

$$(\boldsymbol{g}_k, \boldsymbol{h}_k) = \underset{\substack{\boldsymbol{g} \in \mathbb{R}^r, \boldsymbol{h} \in \mathbb{R}^s \\ \boldsymbol{g}' \Sigma_{11} \boldsymbol{g}_j = 0, \\ \boldsymbol{h}' \Sigma_{22} \boldsymbol{h}_j = 0, j = 1, \ldots, k-1}}{\operatorname{argmax}} \frac{\boldsymbol{g}' \Sigma_{12} \boldsymbol{h}}{\sqrt{\boldsymbol{g}' \Sigma_{11} \boldsymbol{g} \, \boldsymbol{h}' \Sigma_{22} \boldsymbol{h}}}. \tag{6}$$

- $(\boldsymbol{g}_k, \boldsymbol{h}_k)$ Canonical correlation vectors.
- $(\xi_k = \boldsymbol{g}_k' \boldsymbol{X}, \omega_k = \boldsymbol{h}_k' \boldsymbol{Y})$ the kth canonical variates.
- $\rho_k = \text{Corr}(\xi_k, \omega_k) \le \rho_j$, $j = 1, \ldots, k-1$.
- $\text{Corr}(\xi_k, \xi_j) = 0, \text{Corr}(\omega_k, \omega_j) = 0$, $j = 1, \ldots, k-1$.
- Generally $\boldsymbol{g}_i' \boldsymbol{g}_j = 0$ NOT true.

For n sample $(\boldsymbol{x}_1, \boldsymbol{y}_i)$, $(i = 1, \ldots, n)$, denote

$$\mathbf{S}_{11} = \widehat{\mathrm{Cov}}(\boldsymbol{X}), \mathbf{S}_{22} = \widehat{\mathrm{Cov}}(\boldsymbol{Y})$$
$$\mathbf{S}_{12} = \widehat{\mathrm{Cov}}(\boldsymbol{X}, \boldsymbol{Y}), \mathbf{S}_{21} = \widehat{\mathrm{Cov}}(\boldsymbol{Y}, \boldsymbol{X}).$$

The sample CCA is

$$(\boldsymbol{g}_k, \boldsymbol{h}_k) = \underset{\substack{\boldsymbol{g}\in\mathbb{R}^r, \boldsymbol{h}\in\mathbb{R}^s \\ \boldsymbol{g}'\mathbf{S}_{11}\boldsymbol{g}_j = 0, \\ \boldsymbol{h}'\mathbf{S}_{22}\boldsymbol{h}_j = 0, j=1,\ldots,k-1}}{\operatorname{argmax}} \frac{\boldsymbol{g}'\mathbf{S}_{12}\boldsymbol{h}}{\sqrt{\boldsymbol{g}'\mathbf{S}_{11}\boldsymbol{g}\,\boldsymbol{h}'\mathbf{S}_{22}\boldsymbol{h}}}. \tag{7}$$

Finding CCA solution

First CC vectors: maximize $\frac{\boldsymbol{g}'\mathbf{S}_{12}\boldsymbol{h}}{\sqrt{\boldsymbol{g}'\mathbf{S}_{11}\boldsymbol{g}\boldsymbol{h}'\mathbf{S}_{22}\boldsymbol{h}}}$.

Change-of-variable: $\boldsymbol{a} = \mathbf{S}_{11}^{\frac{1}{2}}\boldsymbol{g}$, $\boldsymbol{b} = \mathbf{S}_{22}^{\frac{1}{2}}\boldsymbol{h}$. The problem is now to maximize

$$\frac{\boldsymbol{a}'\mathbf{S}_{11}^{-\frac{1}{2}}\mathbf{S}_{12}\mathbf{S}_{22}^{-\frac{1}{2}}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{a}\boldsymbol{b}'\boldsymbol{b}}}$$

with respect to $\boldsymbol{a} \in \mathbb{R}^r$, $\boldsymbol{b} \in \mathbb{R}^s$, and the solution is given by

$$\boldsymbol{a}_1 = \boldsymbol{v}_1(\mathbf{S}_{11}^{-\frac{1}{2}}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-\frac{1}{2}}),$$
$$\boldsymbol{b}_1 = \boldsymbol{v}_1(\mathbf{S}_{22}^{-\frac{1}{2}}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-\frac{1}{2}}),$$

where $\boldsymbol{v}_i(\boldsymbol{M})$ is the ith eigenvector (corresponding to the ith largest eigenvalue) of symmetric $\boldsymbol{M}$.

Finding CCA solution

For $j = 1, \ldots, \min(s, r)$,

$$\boldsymbol{a}_j = \boldsymbol{v}_j(\mathbf{S}_{11}^{-\frac{1}{2}}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-\frac{1}{2}}),$$
$$\boldsymbol{b}_j = \boldsymbol{v}_j(\mathbf{S}_{22}^{-\frac{1}{2}}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-\frac{1}{2}}),$$

we have

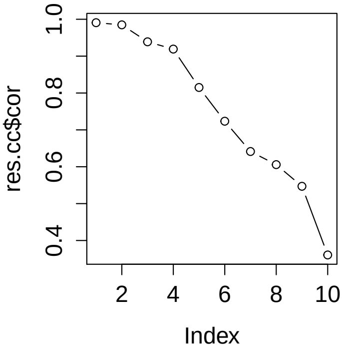
$$\boldsymbol{g}_j = \mathbf{S}_{11}^{-\frac{1}{2}}\boldsymbol{a}_j, \quad \boldsymbol{h}_j = \mathbf{S}_{22}^{-\frac{1}{2}}\boldsymbol{b}_j$$

and thus canonical covariate scores $\boldsymbol{\xi}_j = \boldsymbol{g}_j'\boldsymbol{X} = (\boldsymbol{g}_j'\boldsymbol{x}_1, \ldots, \boldsymbol{g}_j'\boldsymbol{x}_n)$ and $\boldsymbol{\omega}_j = \boldsymbol{h}_j'\boldsymbol{Y} = (\boldsymbol{h}_j'\boldsymbol{y}_1, \ldots, \boldsymbol{h}_j'\boldsymbol{y}_n)$. Note that

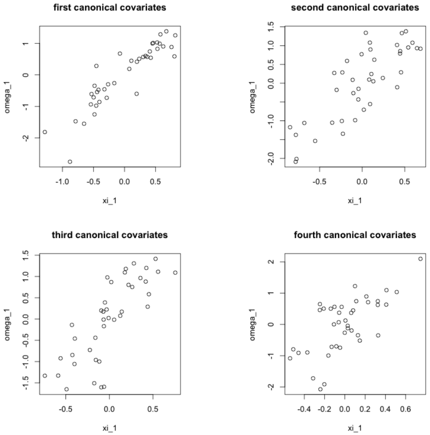
- $\boldsymbol{g}_j'\boldsymbol{g}_k = \boldsymbol{a}_j'\mathbf{S}_{11}^{-1}\boldsymbol{a}_k \neq 0$ (not orthogonal);
- $\widehat{\mathrm{Cov}}(\{\xi_{j(i)}\}, \{\xi_{k(i)}\}) = \boldsymbol{g}_j'\mathbf{S}_{11}\boldsymbol{g}_k = 0$ (uncorrelated).

`nutrimouse` data
Canonical correlation coefficients

- Reduced dataset: $r = 10$ and $s = 21$ variables with $n = 40$ samples.
- $\min(s, r) = 10$ pairs of canonical variates, with decreasing canonical correlation coefficients.

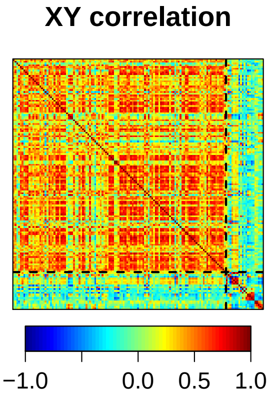


- First four pairs of canonical variates $\{(\xi_{(j)i}, \omega_{(j)i}) : i = 1, \ldots, n\}$.



nutrimouse data with $r = 120$

Original data, with $r \gg n$. Problem?



Regularized CCA

When $r \geq n$, and $s \geq n$, there always exist

$$(\boldsymbol{g}_1, \boldsymbol{h}_1) = \underset{\boldsymbol{g}\in\mathbb{R}^r, \boldsymbol{h}\in\mathbb{R}^s}{\operatorname{argmax}} \frac{\boldsymbol{g}'\mathbf{S}_{12}\boldsymbol{h}}{\sqrt{\boldsymbol{g}'\mathbf{S}_{11}\boldsymbol{g}\,\boldsymbol{h}'\mathbf{S}_{22}\boldsymbol{h}}},$$

satisfying $\mathrm{Corr}(\xi_j, \omega_k) = \pm 1$ (perfect correlation!). This is an artifact from that $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are not invertible (or the rank of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ is at most $n-1$.)

A way to circumvent this issue and obtain a meaningful estimate of population canonical correlation coefficients, Regularized CCA is often used:

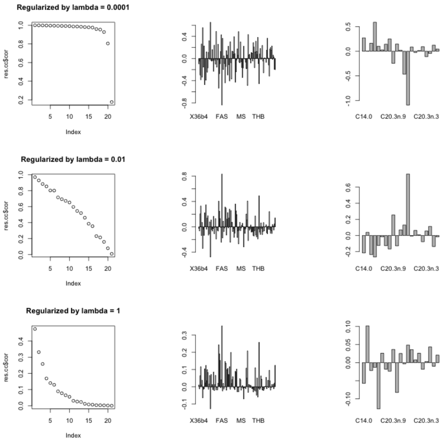
$$(\boldsymbol{g}_k, \boldsymbol{h}_k) = \underset{\substack{\boldsymbol{g}\in\mathbb{R}^r, \boldsymbol{h}\in\mathbb{R}^s \\ \boldsymbol{g}'\mathbf{S}_{11}\boldsymbol{g}_j=0, \\ \boldsymbol{h}'\mathbf{S}_{22}\boldsymbol{h}_j=0, j=1,\ldots,k-1}}{\operatorname{argmax}} \frac{\boldsymbol{g}'\mathbf{S}_{12}\boldsymbol{h}}{\sqrt{\boldsymbol{g}'(\mathbf{S}_{11}+\lambda_1\mathbb{I}_r)\boldsymbol{g}\,\boldsymbol{h}'(\mathbf{S}_{22}+\lambda_2\mathbb{I}_s)\boldsymbol{h}}}, \tag{8}$$

for $\lambda_1, \lambda_2 > 0$.

nutrimouse data–Regularized CCA

1. sample canonical correlation coefficients ρ_j and
2. corresponding projection vectors $\boldsymbol{h}_1$ and $\boldsymbol{g}_1$

are varing depending on the choice of regularization parameter $\lambda_1 = \lambda_2 = 0.0001, 0.01, 1$.



CCA in R

$n \times r$ data frame or matrix `x` and $n \times s$ data frame or matrix `y`
To perform CCA:

```
cc=cancor(x,y)
rho<-cc$cor
g<-cc$xcoef
h<-cc$ycoef
```

To perform regularized CCA, use package `CCA`,

```
library(CCA)
cc=rcc(x,y,lambda1,lambda2)
rho<-cc$cor
g<-cc$xcoef
h<-cc$ycoef
```

Possible research topics

Sparse PCA

- H Shen, JZ Huang - Journal of multivariate analysis, 2008
- A d'Aspremont, L El Ghaoui, MI Jordan, GRG Lanckriet - SIAM review, 2007
- Z Ma - The Annals of Statistics, 2013
- TT Cai, Z Ma, Y Wu - The Annals of Statistics, 2013

PCA Consistency

- S Jung, JS Marron - The Annals of Statistics, 2009
- D Shen, H Shen, JS Marron - Journal of Multivariate Analysis, 2013
- S Jung, A Sen, JS Marron - Journal of Multivariate Analysis, 2012