Lecture 13: Principal Component Analysis

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

Read: ELSII Ch. 14.5, ISLR 10.2 & 10.4, and SLS 8.2

Outline

From Supervised Learning to Unsupervised Learning

- Supervised Learning: regression and classification. Response Y is used in training and the goal is to predict Y
- Unsupervised Learning: No information about the response is used. The goal is to understand the X data.

The next section would be . . . . . .

The next section would be . . . . . .

Data Visualization

- Suppose there are variables $X_1, \ldots, X_p$. To visualize the data, one may draw pairwise scatterplots. But there are $p(p-1)/2$ such plots.
- Data lie in p-dimensional space, but not all the dimensions are interesting.
- Solution: find a low-dimensional representation of the data that captures as much of the information as possible.
- PCA seeks a small number of dimensions that are **as interesting as possible**, where interesting-ness is measured by the amount that the observations **vary along each dimension**.

Data Exploration - Multivariate data

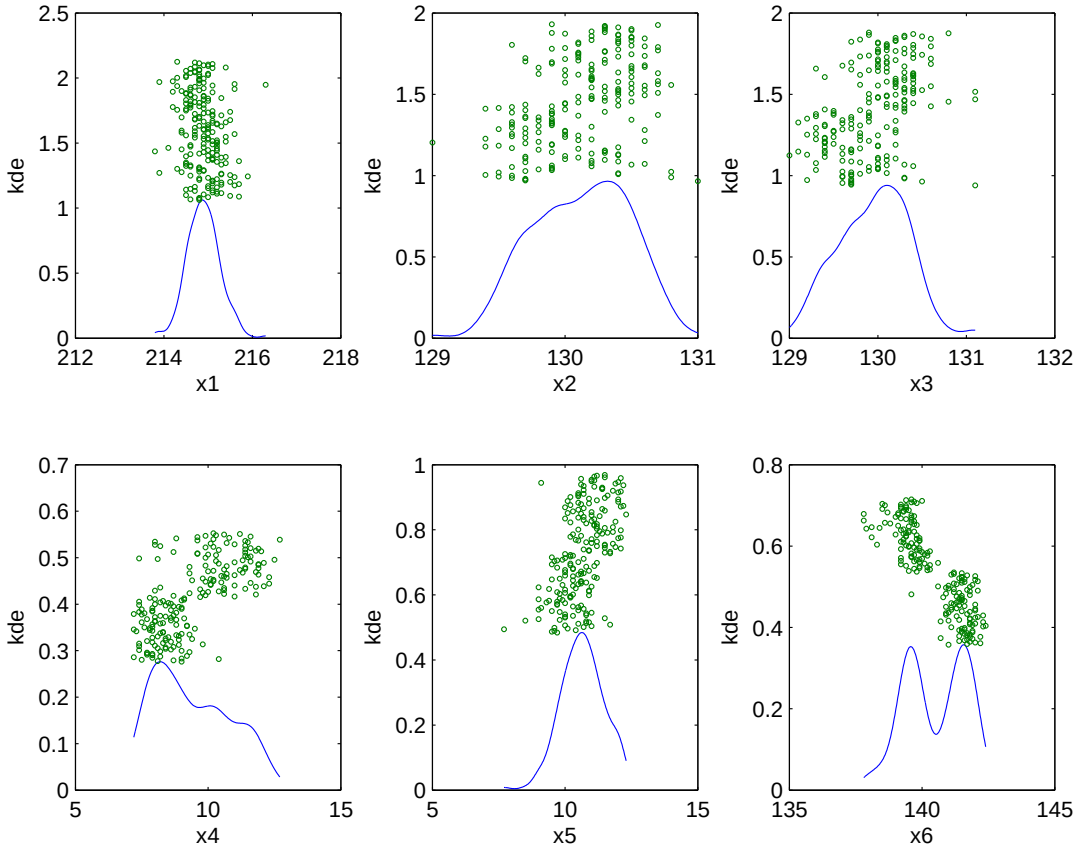
Dimension $p = 6$ example – Swiss bank notes

- $n = 200$ Swiss bank notes (See Fig. 1.1, Härdle and Simar)
- Each note (obs.) has $p = 6$ measurements (variables).
- Additional information: first half are genuine; the other half are counterfeit.
- Visualization of 6-dim'l data?
- Can use 6 KDEs overlaid with jitterplot for each measurements (variables)
- jitterplot: heights of dots (y value) are random for visualization. The x value represents the realized value of the data point.

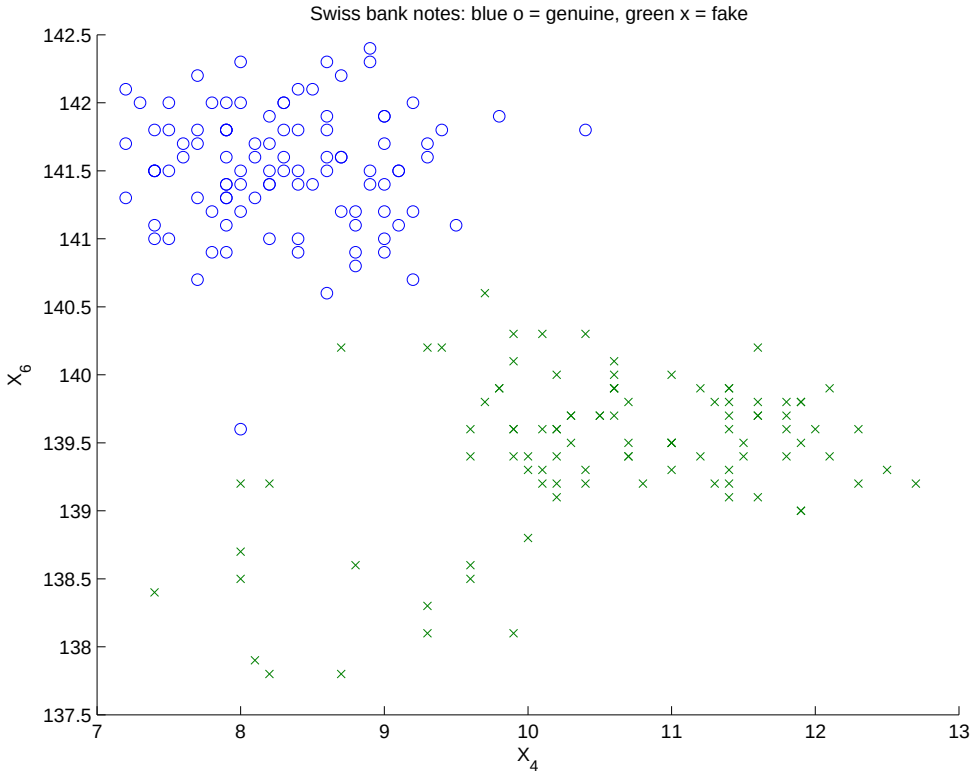
Swiss bank notes - Marginal KDEs

Marginal KDEs overlaid with jitterplot for each of 6 variables.

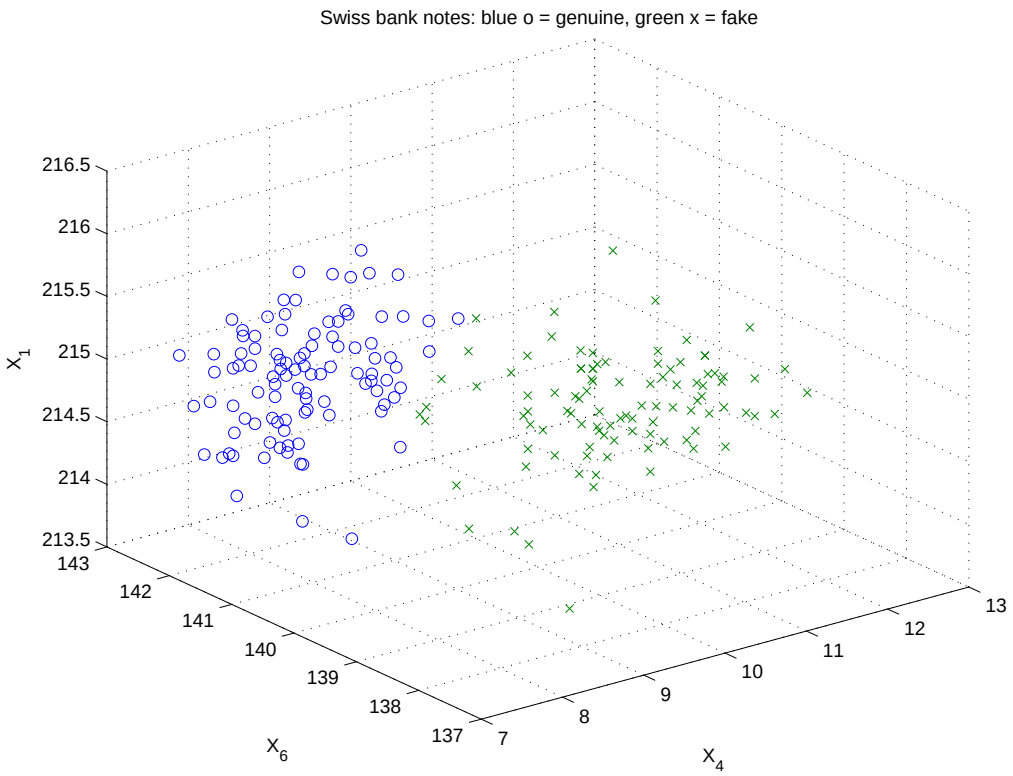
- Informative, realistic when p is small
- No information about association between variables.



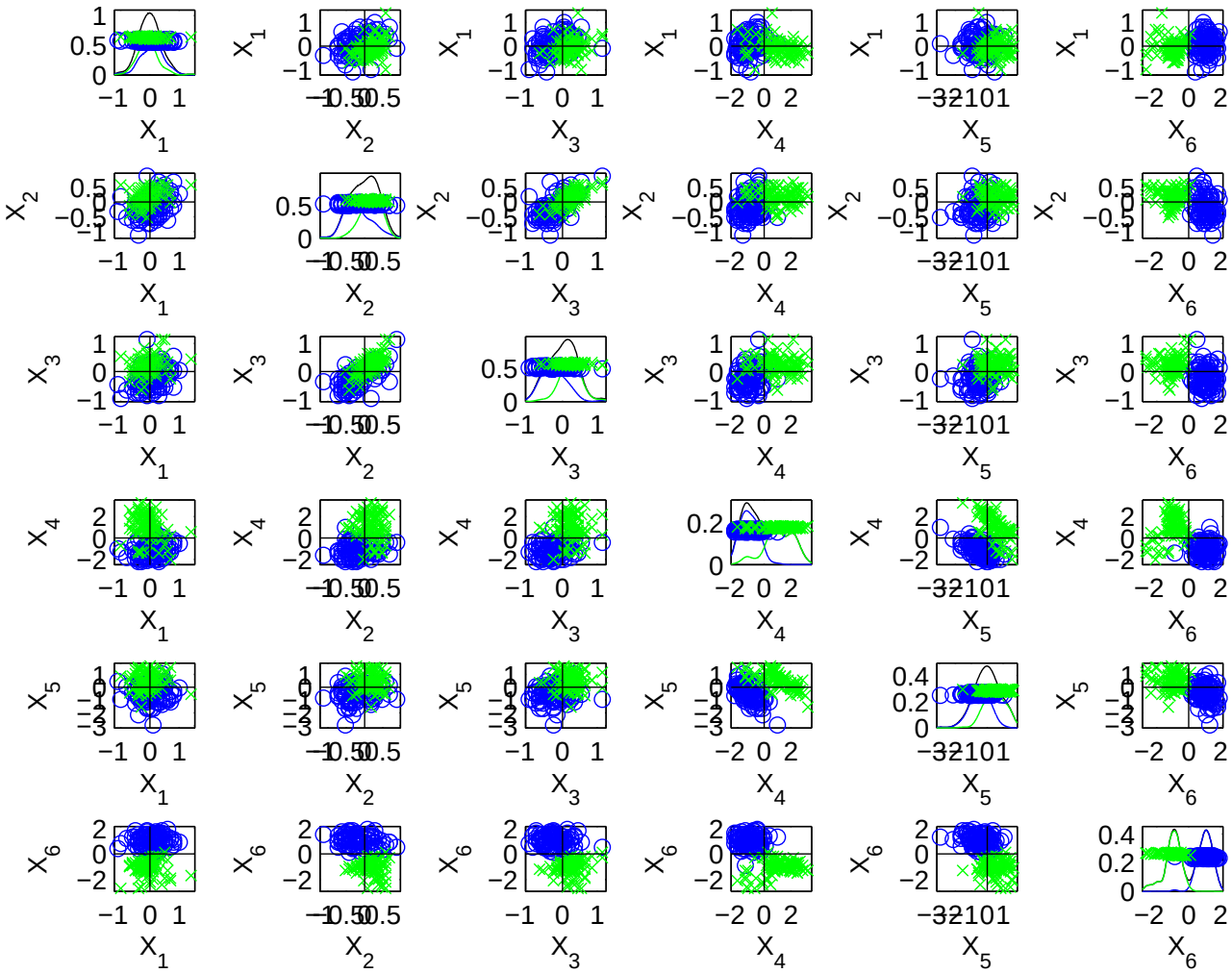
- Variable pair best visualized by scatterplot, e.g. X_4 vs X_6.
- Understood as point clouds, which empirically representing the distribution
- $\binom{6}{2} = 15$ many pairs to choose from. 15 plots?



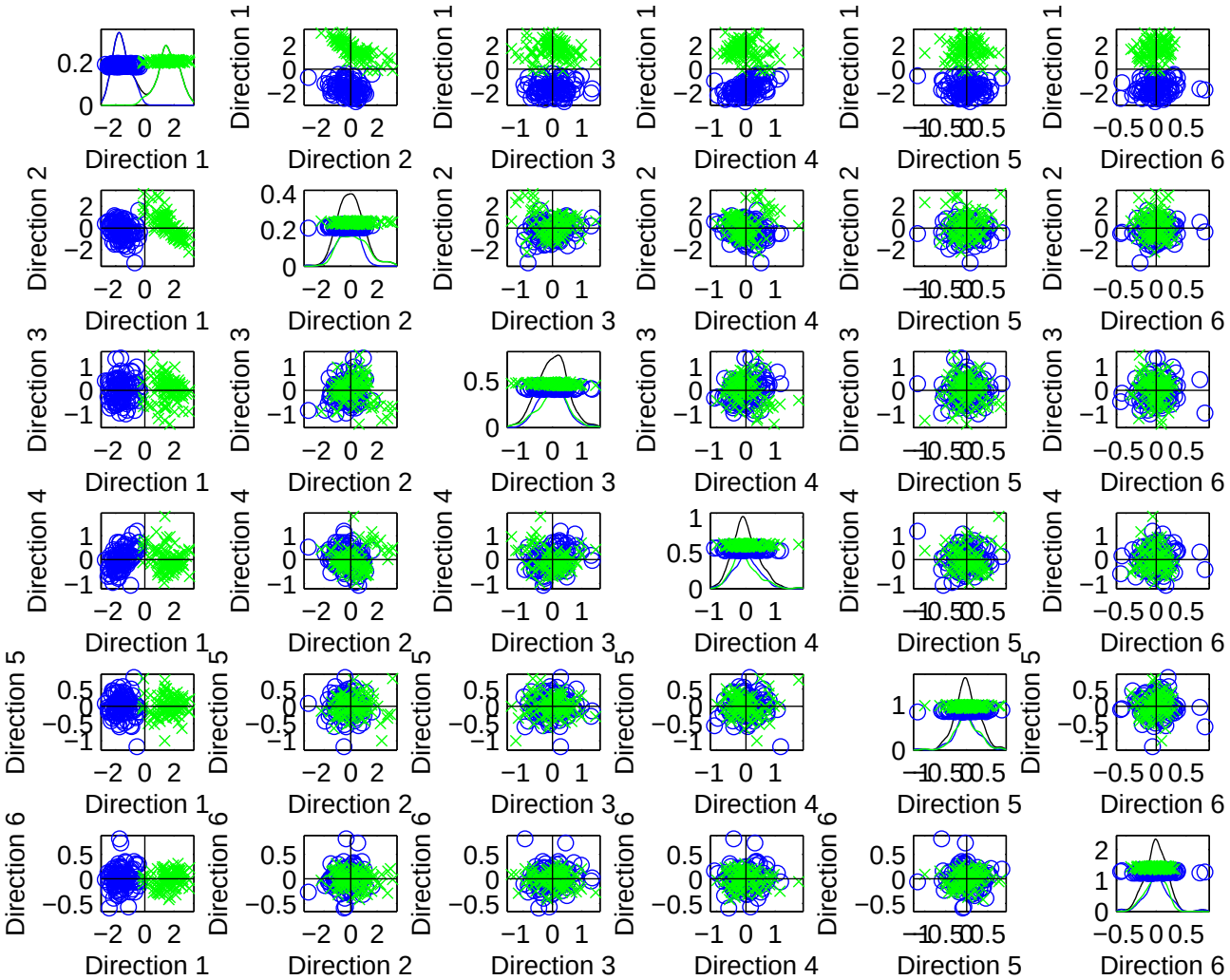
- Scatters of three variables can also be informative
- But only if software allows to rotate the axes.
- Otherwise, the 3D scatterplot is just a 2D scatterplot of two linear combinations of the three variables.
- Angle matters.



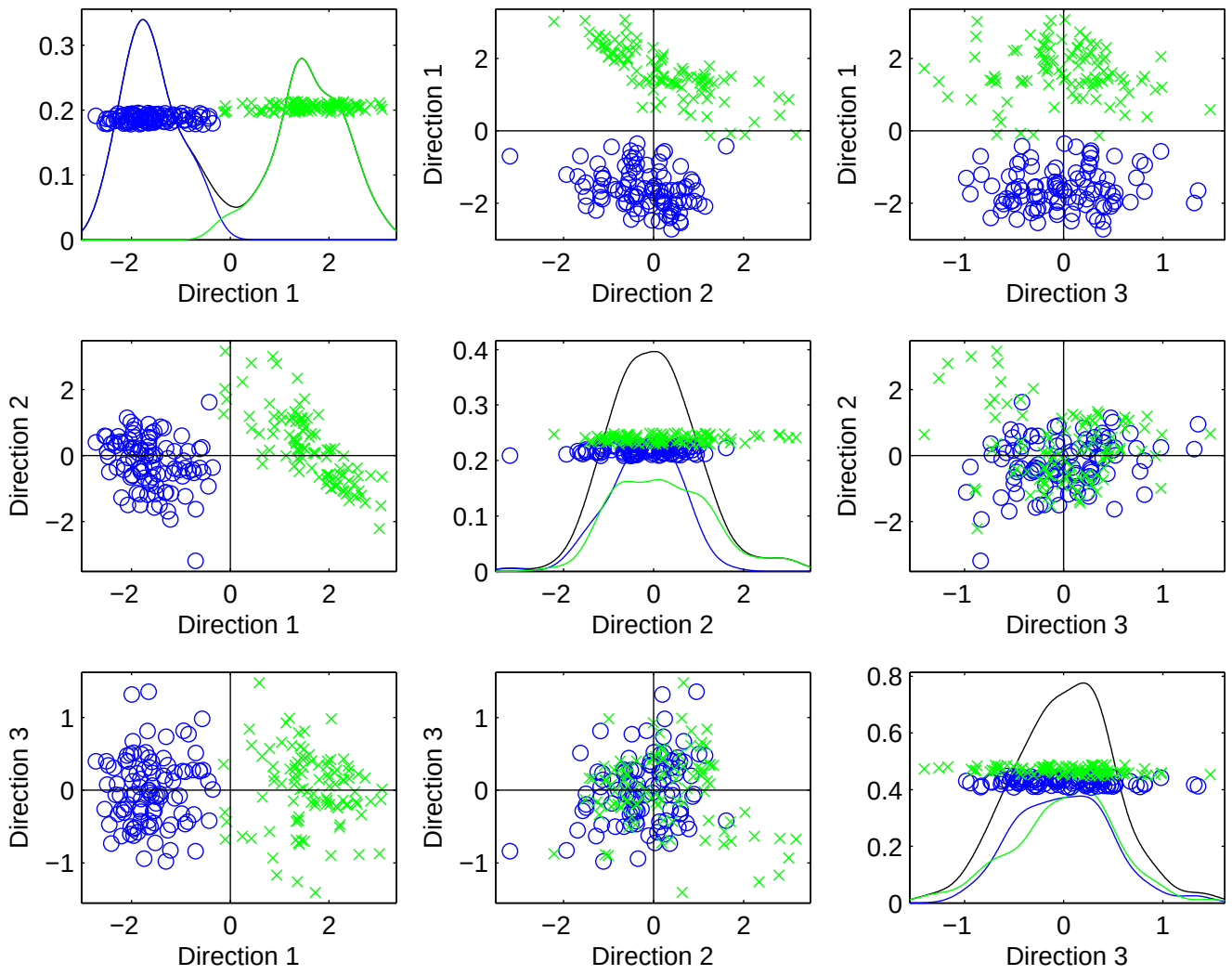
- A traditional, yet powerful, tool is to construct a matrix of scatterplots. - Too busy with $p = 6$.



- Better to visualize with principal component scores.



- With principal component scores, we can focus on **fewer** combinations (it's called Dimension Reduction)



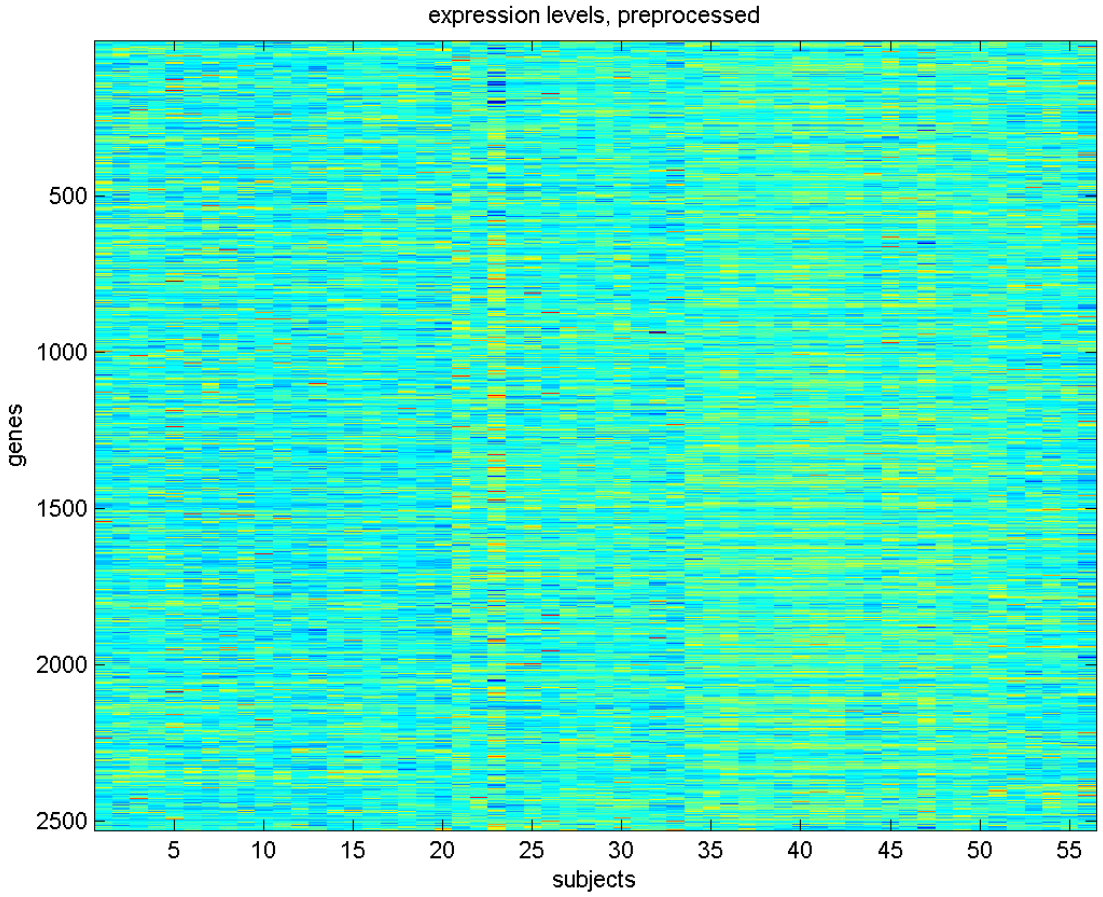
The next section would be . . . . . .

Pattern Recognition

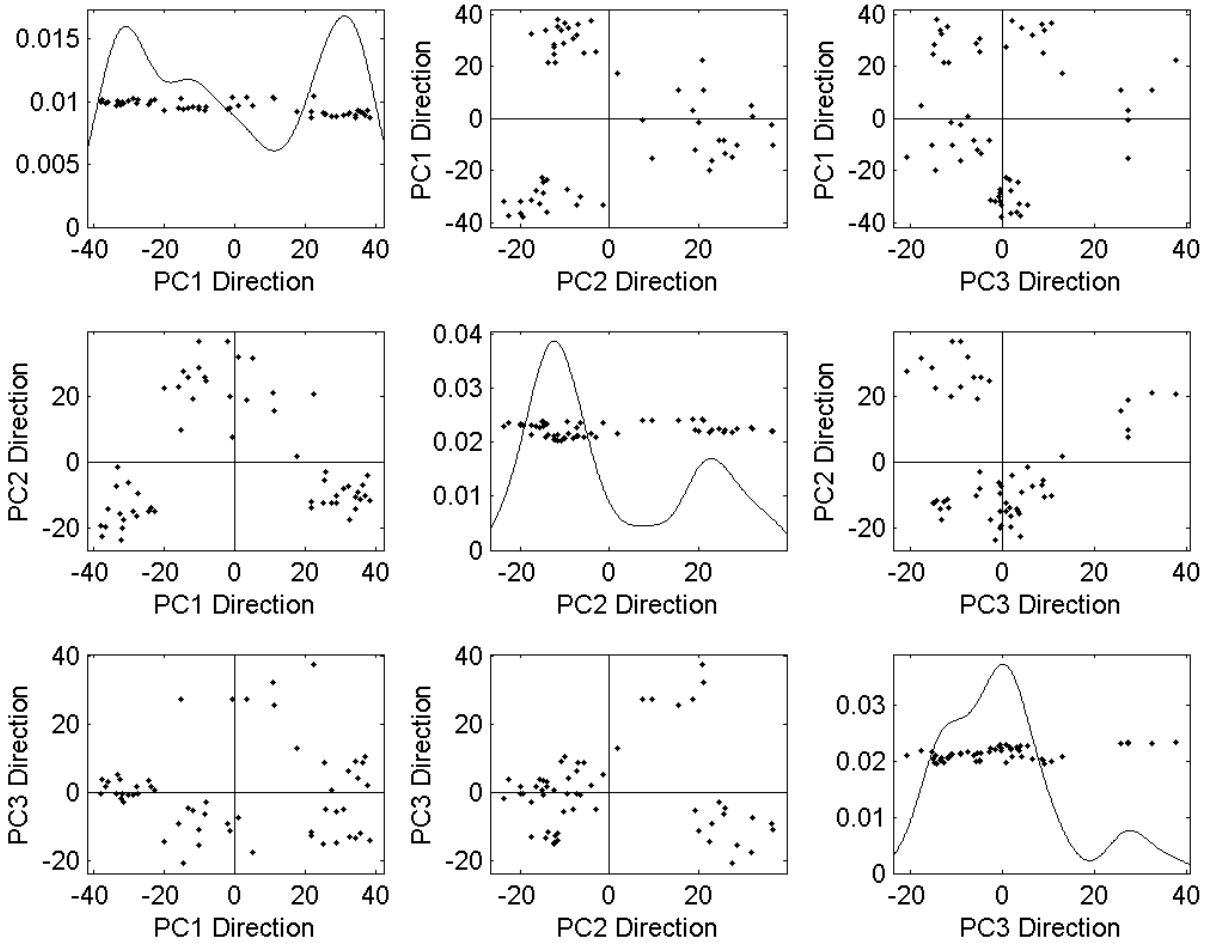
PCA can sometimes help discover previously unknown patterns, and help learn from labelled data.

Example: mRNA expression profiling

- Bhattacharjee et al (2001) PNAS
- Preprocessed gene expressions with $d = 2530$ genes and $n = 56$ subjects with lung cancer.
- Subgroup for different types of lung cancers?

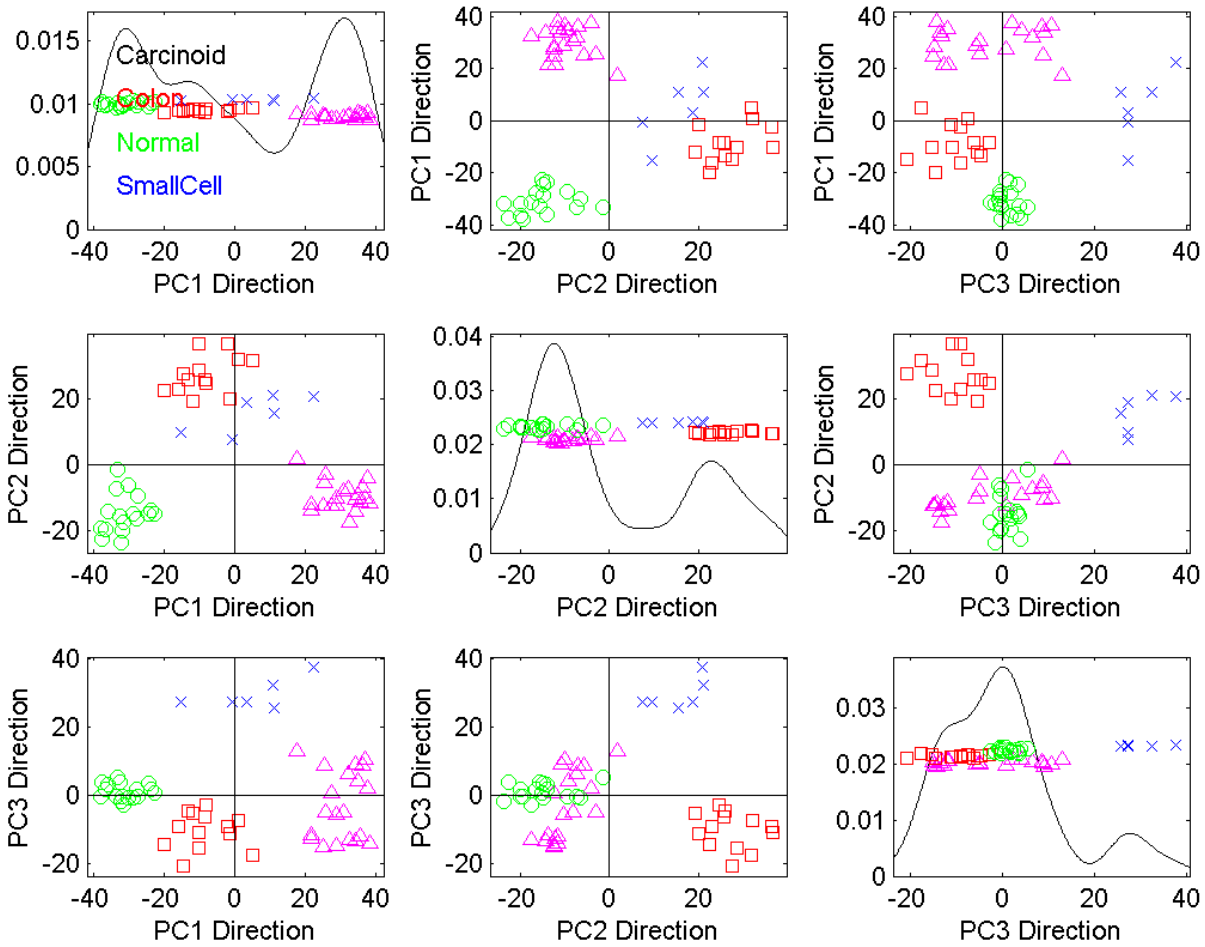


mRNA expression profiling



mRNA expression profiling

- Color by true subgroups



Successfully capture the major pattern in the data: **the black, red, green, blue and cyan observations that are near each other in the high-dimensional space remain nearby in these two-dimensional representations.**

The next section would be . . . . . .

Dimension Reduction

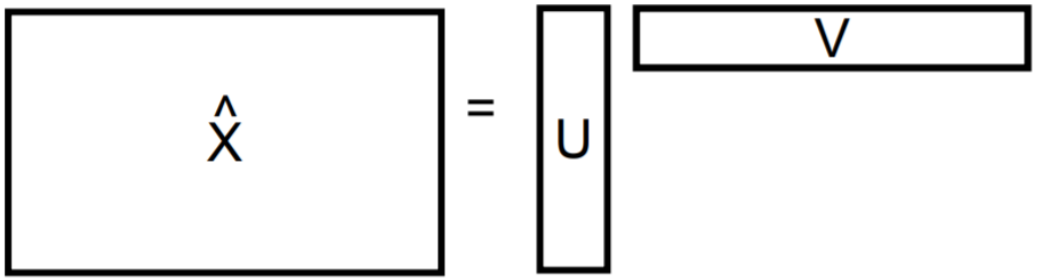
- Lastly, PCA is a way of dimension reduction.
- For example, as in principal component regression, we simply use principal components as predictors in a regression model in place of the original larger set of variables.

The next section would be . . . . . .

The next section would be . . . . . .

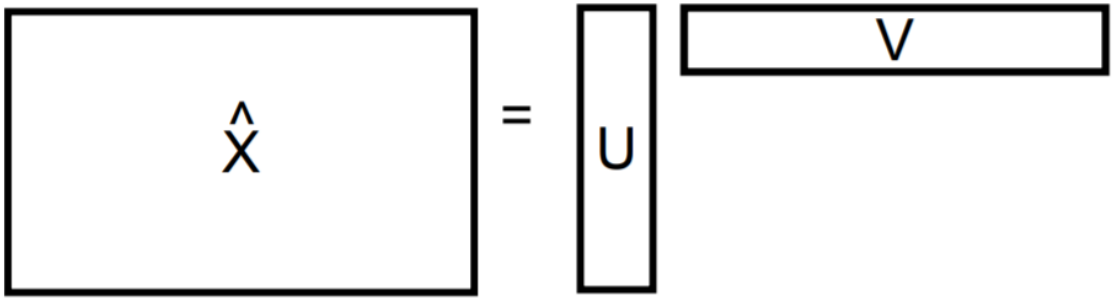
Matrix factorization

- Given matrix X, we seek to find matrices U and V such that $X \approx UV := \hat{X}$.



- It is ideal that U has very few columns.
- Why are low-rank approximations important?
  - Intuitively, if matrix is low rank, then the observations can be explained by linear combinations of **few** underlying factors
  - Want to know which factors control the observations

PCA



- Imagine that X is the **centered** data matrix where the ith column $X_{(i)}$ is the ith observation.
- $X \approx UV$ means that we seek to find U and V so that $X_{(i)} \approx \sum_{j=1}^{q} v_{ji} U_j$ where U_j is the jth column of U.

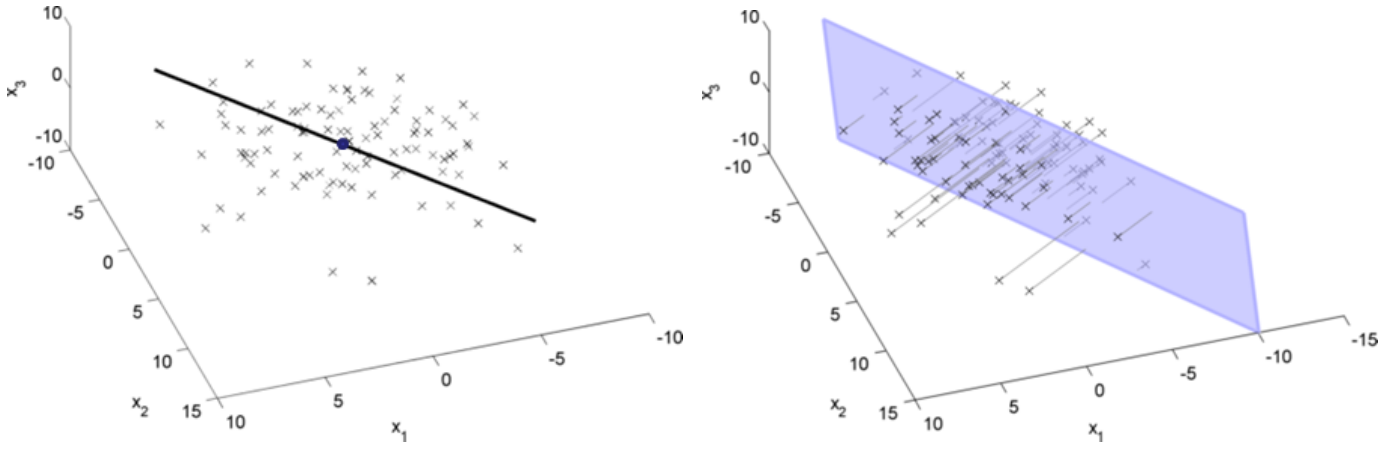
To resolve identifiable issue, we may require $U^T U = \mathbb{I}$.

PCA can be viewed as the following matrix factorization / matrix approximation problem.

$$(U, V) = \underset{U,V,U^T U=\mathbb{I}}{\operatorname{argmin}} \|X - UV\|_F^2 = \underset{U,V,U^T U=\mathbb{I}}{\operatorname{argmin}} \sum_{i=1}^{n} \|X_{(i)} - \sum_{j=1}^{q} v_{ji} U_j\|_2^2$$

where $\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2$ is the Frobenius norm of matrix A

Geometric understanding of PCA



A 3D point cloud.
Mean is removed
left: best 1-d approximation ($q = 1$)
right: best 2-d approximation ($q = 2$)

Next, another formulation of PCA using eigen-decompostion of covariance matrix.

The next section would be . . . . . .

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_i's are most important among all choices.

Principal Component Analysis

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.

PCA aims for a set of direction vectors which lead to maximal variances of the projected random variables.

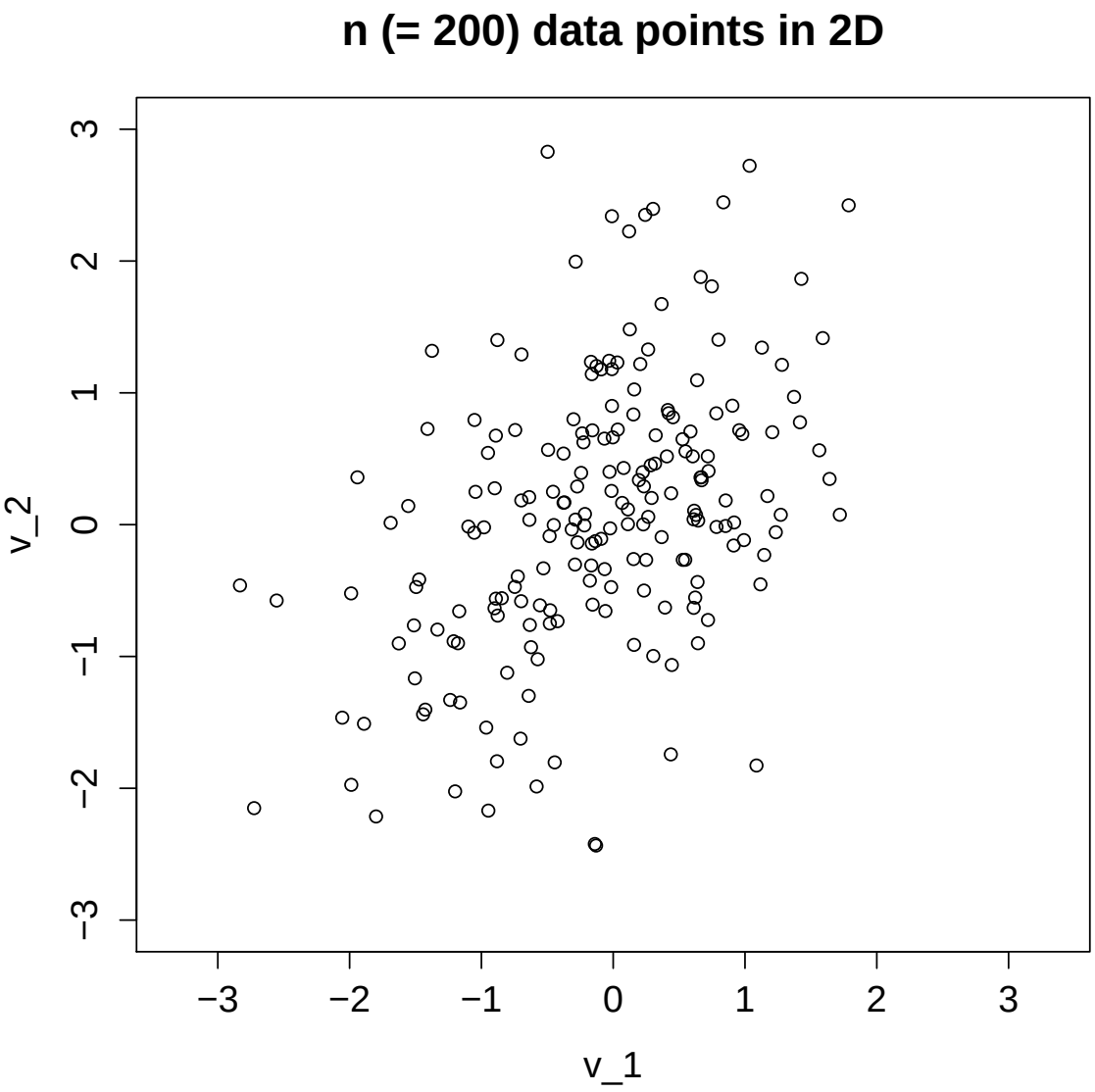
Take $d = 1$. PCA for the distribution of $\boldsymbol{X}$ finds $\boldsymbol{a}_1$ such that

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

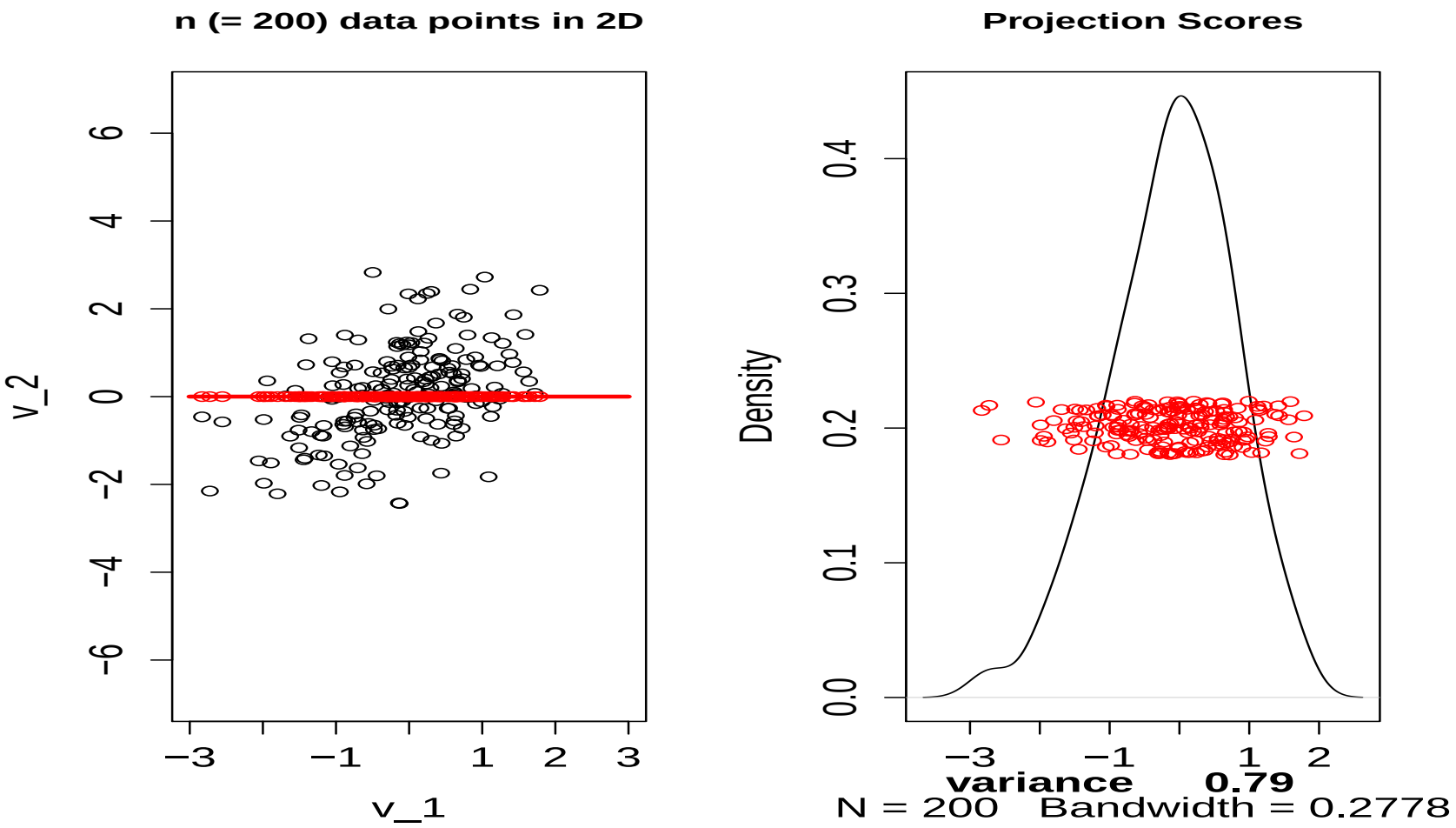
where $Z_1(\boldsymbol{a}) = a_1 X_1 + \cdots + a_p X_p = \boldsymbol{a}' \boldsymbol{X}$.

Geometric understanding of PCA for point cloud

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

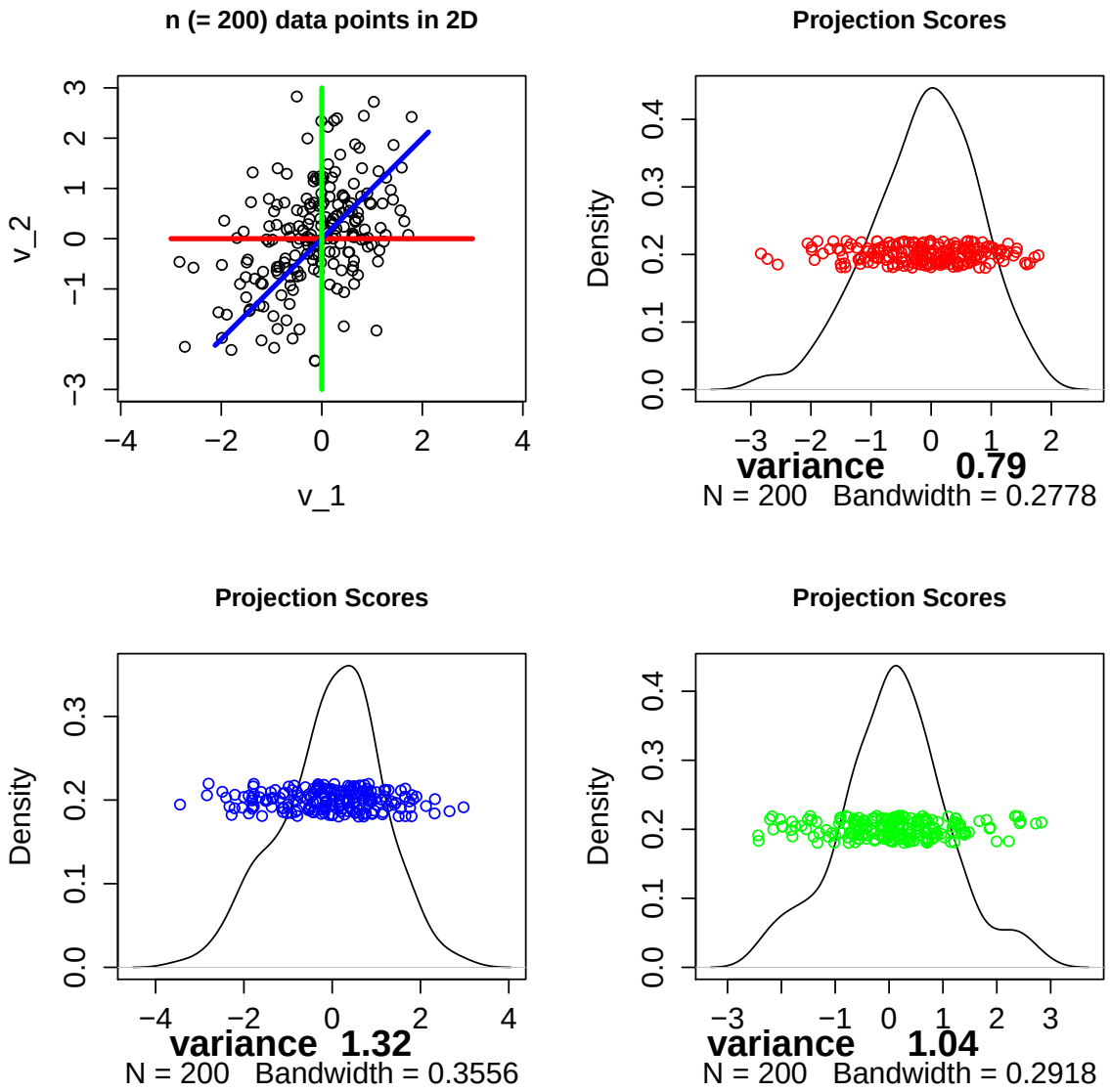


Take $\boldsymbol{a} = (1, 0)'$.



Which one is better?

Take $\boldsymbol{a} = (1,0)', (0,1)', (1/\sqrt{2}, 1/\sqrt{2})'$.



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 $\boldsymbol{u}_1'\boldsymbol{X}$ among all 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 (it's a 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

- can maximize the variance of $\boldsymbol{u}_2'\boldsymbol{X}$;
- among directions orthogonal to the first PC direction $\boldsymbol{u}_1$.

That is,

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

- $\boldsymbol{u}_2 = (u_{21}, \ldots, u_{2p})'$ 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.
- $\lambda_2 = \operatorname{Var}(Z_2)$ is the variance explained by the second PC, and $\lambda_1 \geq \lambda_2$.
- $\operatorname{Corr}(Z_1, Z_2) = 0$.

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 $\boldsymbol{u}_k'\boldsymbol{X}$;
- among those orthogonal to the 1st to the $(k-1)$th PC directions $\boldsymbol{u}_j$ $(j = 1, \ldots, k-1)$

That is,

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

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

Relation to eigen-decomposition of Σ

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

- $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p]$ orthogonal matrix
- $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ 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

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

Gradient

Let $f : \mathbb{R}^d \to \mathbb{R}$. Define

$$\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_d} \end{pmatrix}_{d \times 1}$$

Facts:

- $$\nabla_{\boldsymbol{x}}(\boldsymbol{c}'\boldsymbol{x}) = \frac{\partial(\boldsymbol{c}'\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{c}$$
- $$\nabla_{\boldsymbol{x}}(\boldsymbol{x}'\mathbf{A}\boldsymbol{x}) = \frac{\partial(\boldsymbol{x}'\mathbf{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\mathbf{A}$$

  for symmetric $\mathbf{A}$

Relation to eigen-decomposition of Σ

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

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

The stationary point solves the following:

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

which leads to

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

Recall: 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 $\lambda_1 = \boldsymbol{u}_1'\Sigma\boldsymbol{u}_1 = \mathrm{Var}(\boldsymbol{u}_1'\boldsymbol{X}) \geq \lambda_j$ $(j > 1)$.

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

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

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 $\lambda_k = \boldsymbol{u}_k'\Sigma\boldsymbol{u}_k$ among the remaining eigen-pairs.

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 $\tilde{\mathbf{X}}$.

Eigen-decomposition of $\mathbf{S}$

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

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

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

The next section would be . . . . . .

The next section would be . . . . . .

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

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

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

The next section would be . . . . . .

SVD and Eigen-decomposition Connection

If SVD of $\tilde{\mathbf{X}} = \mathbf{UDV}'$, then

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

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

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

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
```

Scaling? Correlation PCA

- PCA is not scale invariant.
- SOMETIMES, good idea to do normalization.
- Correlation matrix of a random vector $\mathbf{X}$ is given by

$$\mathbf{R} = \mathbf{D}_{\Sigma}^{-\frac{1}{2}} \Sigma \mathbf{D}_{\Sigma}^{-\frac{1}{2}},$$

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

The next section would be . . . . . .

PC loadings and PC scores

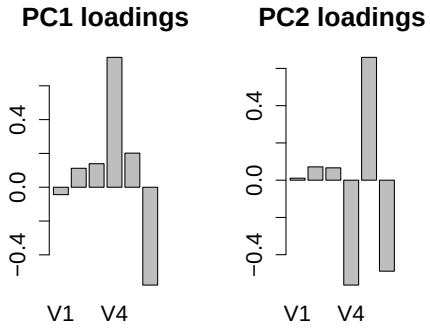
- For some students, it is often confusing between PC loadings and PC scores.
- PC direction $\boldsymbol{U}_j$: jth column of $\mathbf{U}$, a p-dimensional vector.
- PC loadings: elements of $\boldsymbol{U}_j$, measuring contributions from different dimensions (variables) to the jth principal component
- PC scores: inner products of $\boldsymbol{x}_i^T \boldsymbol{U}_j$, $i = 1, \ldots, n$, coordinates of obs. i in the new coordinate system spanned by $\boldsymbol{U}_j$'s

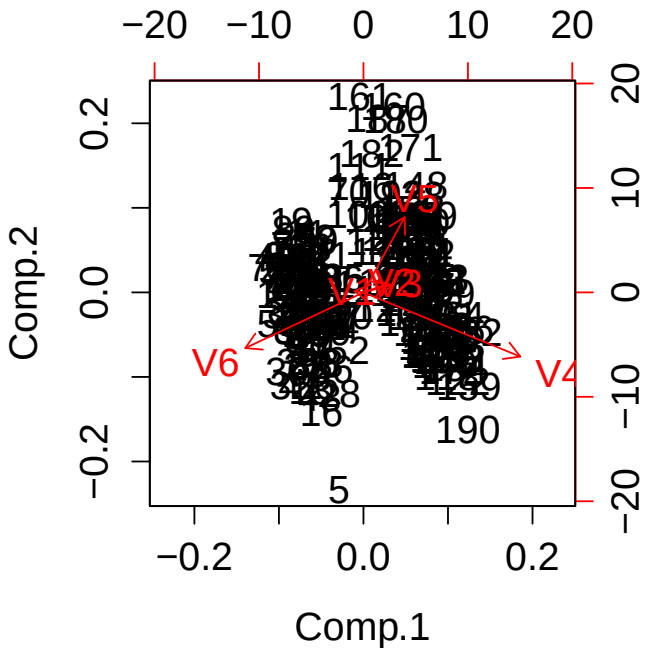
Which variables are most responsible for the principal components?

- Check loadings of principal component directions.
- **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
```





Recall that the scatter plot of PC1+PC2 is a visualization after rotation and projection.

Hence red vectors can be viewed as the rotated and projected coordinate direction vectors. For example, $V1$ is the rotated and projected $(1, 0, 0, 0, ...)'$

The next section would be . . . . . .

How many components to keep? (1)

"How much of the variation within the data have PCs explained?"

1. Total variation in $\mathbf{X}$ is the sum of all marginal (sample) variances

$$\sum_{k=1}^{p} \mathrm{Var}(\{x_{ki} : i = 1, \ldots, n\}) = \mathrm{Trace}(\mathbf{S}) = \mathrm{Trace}(\hat{\Lambda})$$
$$= \sum_{k=1}^{p} \hat{\lambda}_k = \sum_{k=1}^{p} \mathrm{Var}(\{z_{(k)i} : i = 1, \ldots, n\}).$$

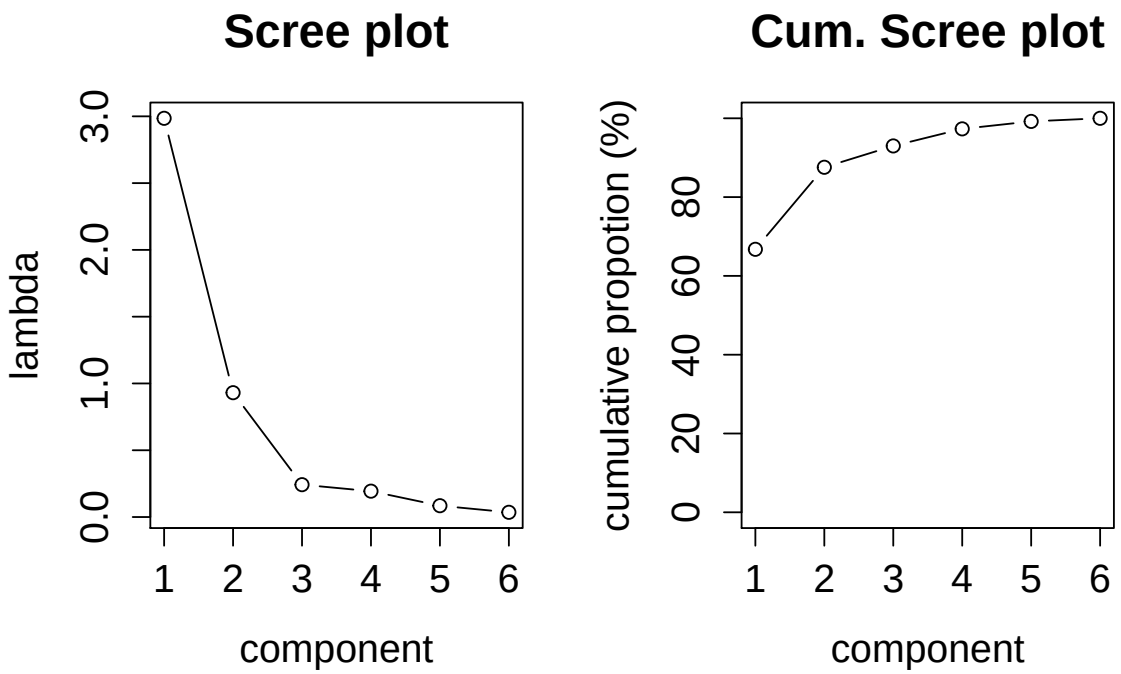
2. (Sample) variance of the kth PC:

$$\mathrm{Var}(\{z_{(k)i} : i = 1, \ldots, n\}) = \hat{\lambda}_k$$

3. Total variance in the 1st to the kth PCs: $\hat{\lambda}_1 + \ldots + \hat{\lambda}_k$.

How many components to keep? (1)

1. In scree plot $(k, \hat{\lambda}_k)$, we look for an elbow.
2. In cumulative scree plot (proportion of variance explained, $(k, \frac{\sum_{j=1}^{k} \hat{\lambda}_j}{\sum_{j=1}^{p} \hat{\lambda}_j})$), 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

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

   The first k components will be retained if $H_0(k)$ is not rejected at a specified level.

The next section would be . . . . . .

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.

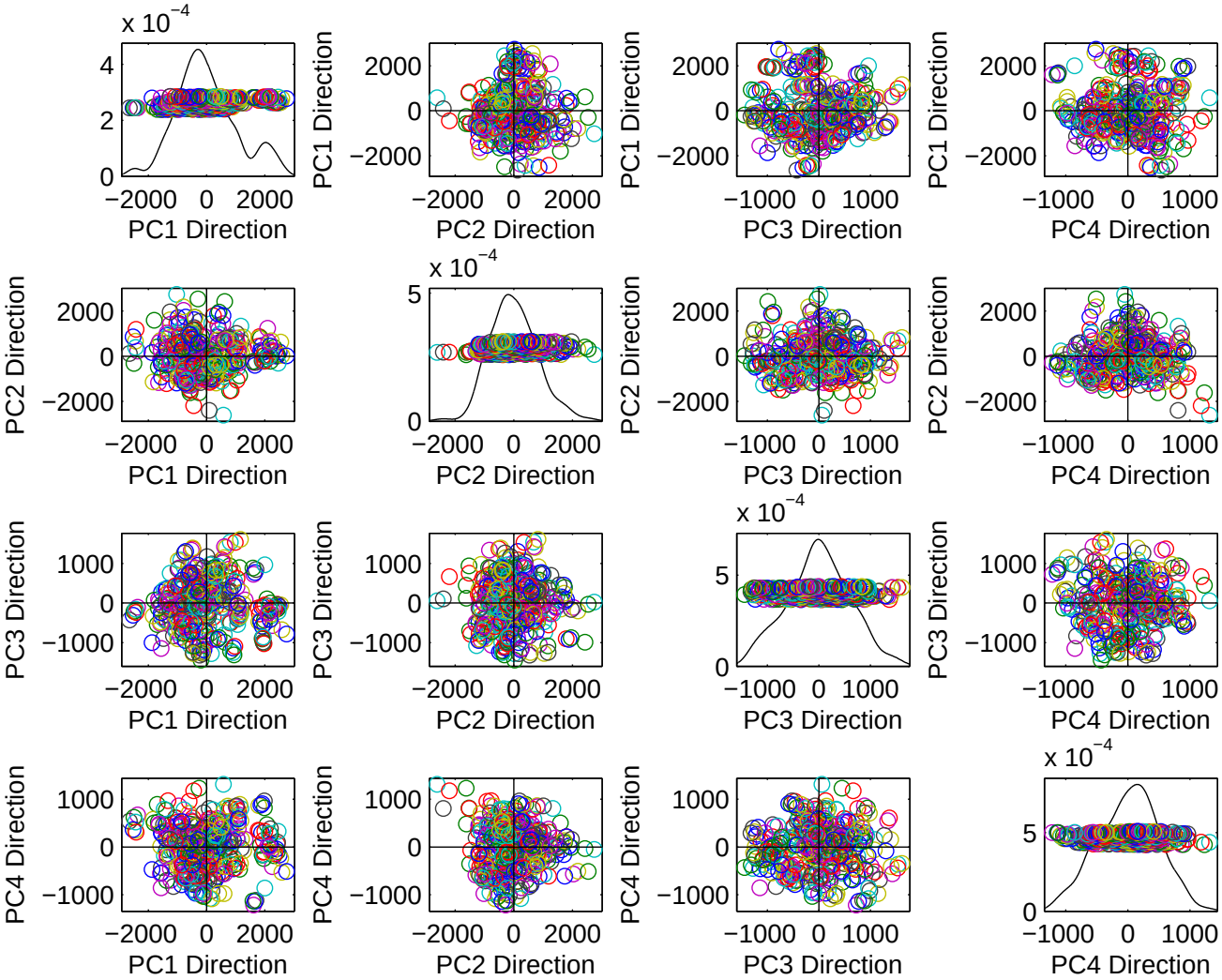
Images as data

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 $d = 4096 = 64 \times 64$.

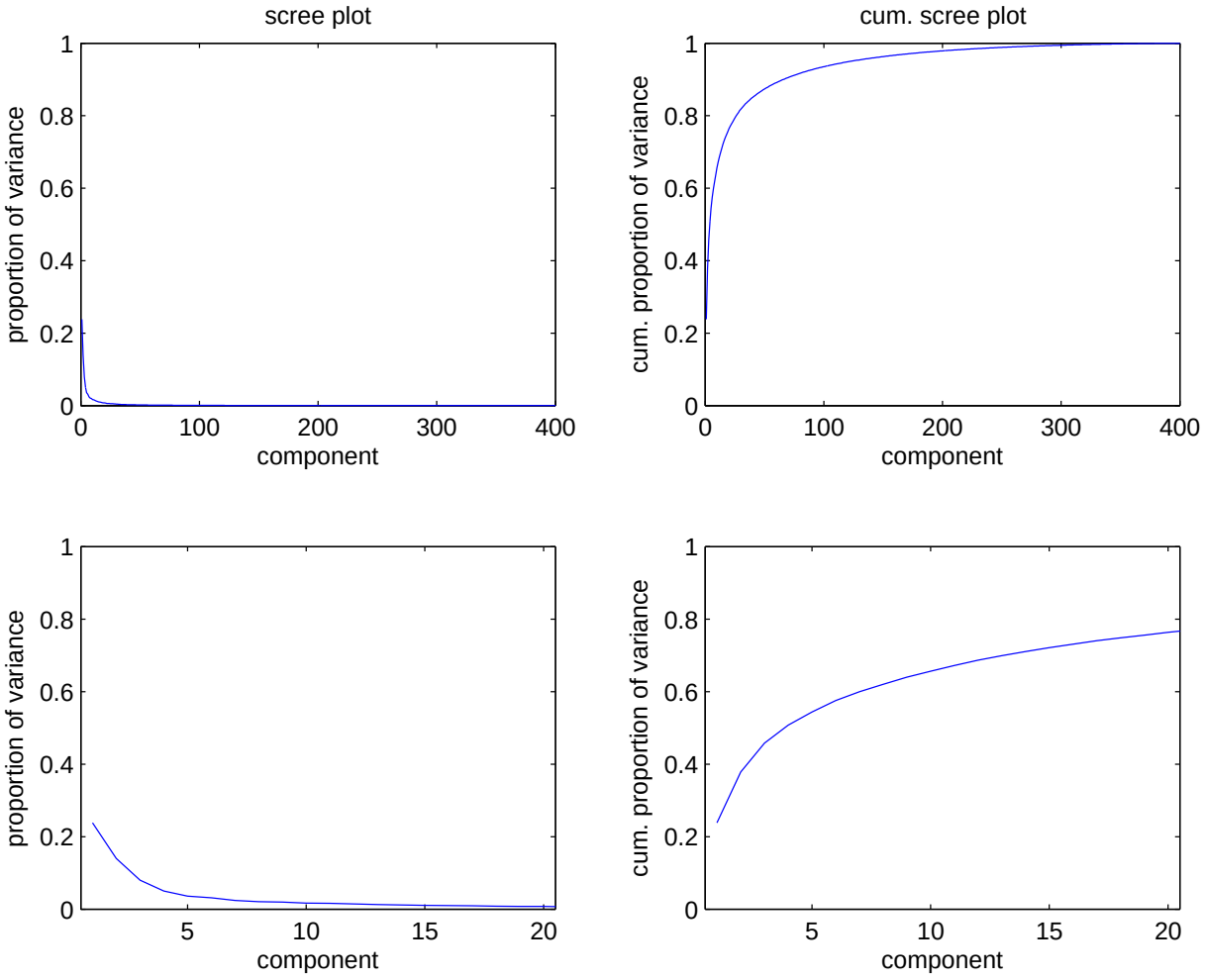
So, $\boldsymbol{x}_1$ is a $d \times 1$ vector corresponding to

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

Olivetti Faces data–Major components

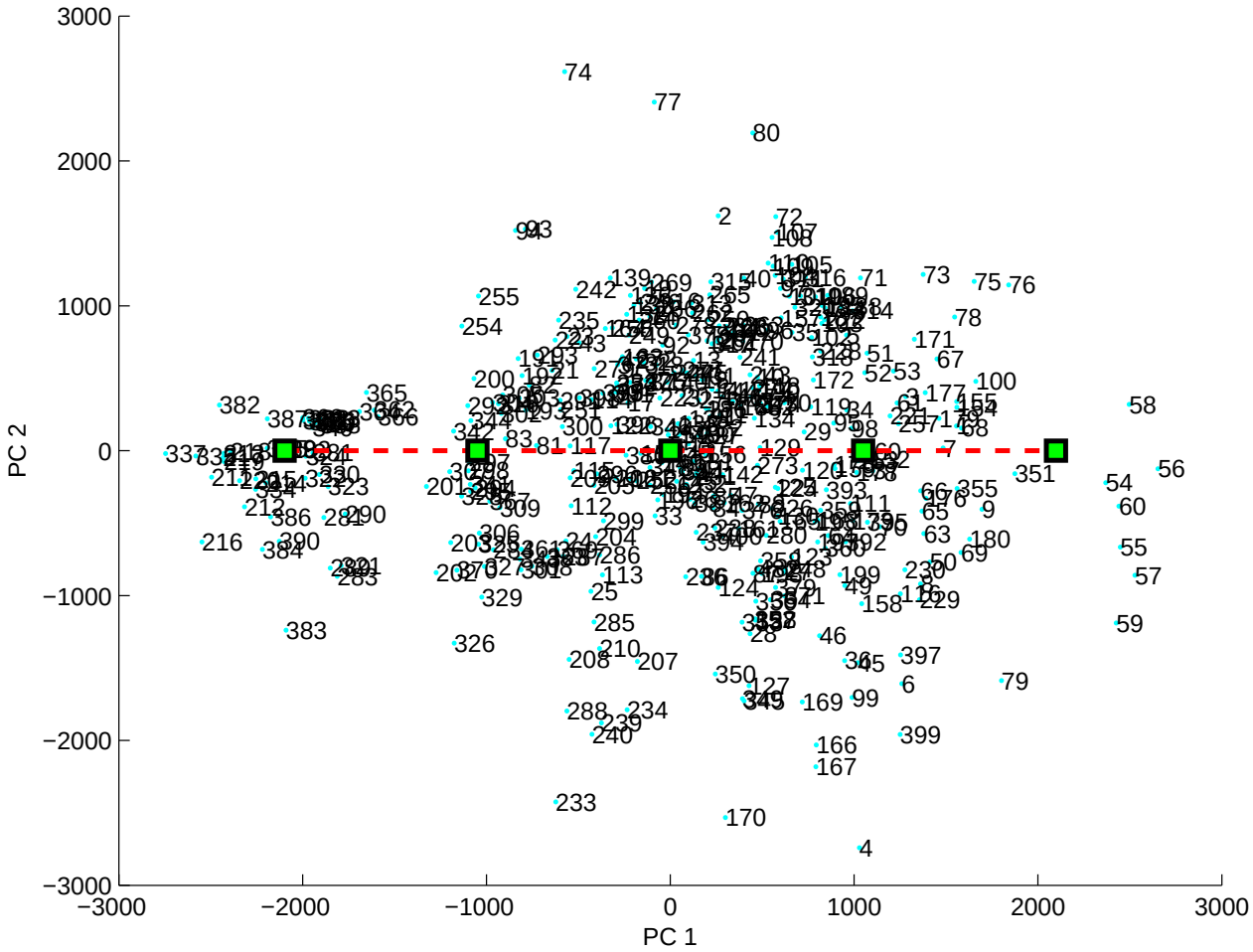


Olivetti Faces data–Scree plots



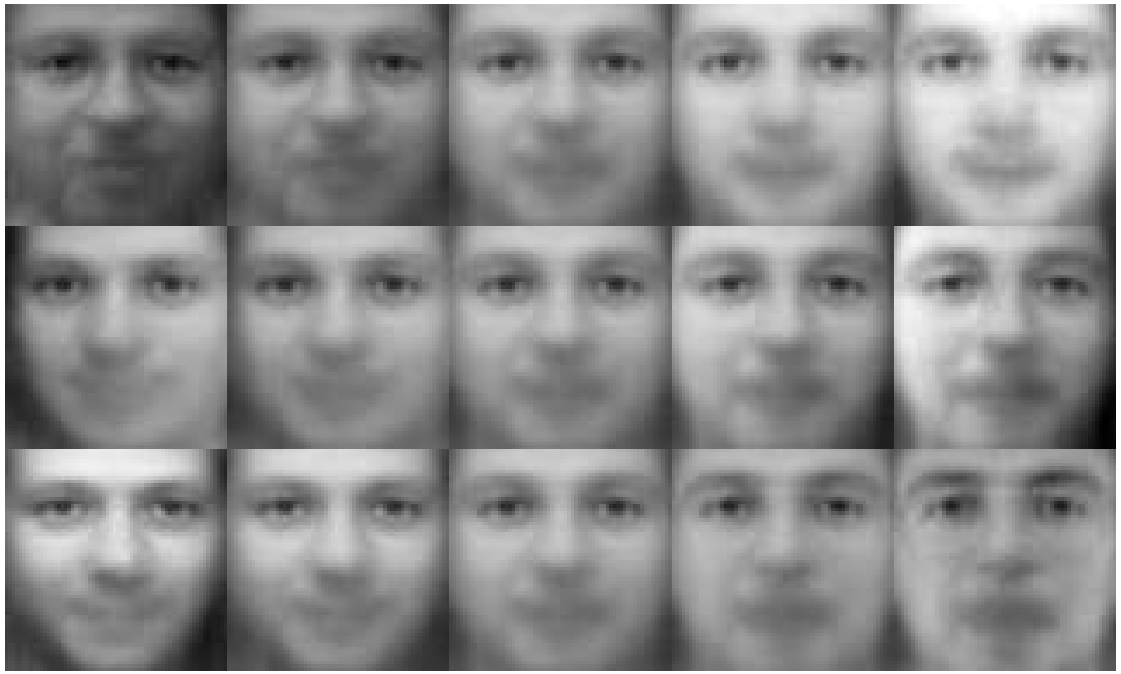
Olivetti Faces data–Interpretation

Examine the mode of variation by walking along the PC direction through the mean. Shown here are $\pm 1, 2$ standard deviations of $Z_{(1)}$ apart from the mean in the direction of PC1.



Olivetti Faces data–Interpretation (Eigenfaces)

PC walk. +−2 std along PC dir. (Top−−PC 1, Mid−−PC 2, Bottom−−PC 3)



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

How to walk along the PC direction?

- Vectorized data in $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, with mean $\bar{\boldsymbol{x}}$.
- Compute $(\boldsymbol{u}_j, \lambda_j)$: 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 $\mathbf{W}_s$.

Next, reconstruct the original face using PCs.

Approximation to the original data matrix

Recall the matrix factorization viewpoint of PCA: $X \approx UV$.

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

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. $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}'$
2. $\mathbf{Z} = \mathbf{U}'\tilde{\mathbf{X}} = \mathbf{D}\mathbf{V}'$
3. $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{Z}$
4. $\tilde{\boldsymbol{x}}_i = \mathbf{U}\boldsymbol{z}_i = \sum_{j=1}^{p} \boldsymbol{u}_j z_{(j)i}$
   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
   $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$.

Hence

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

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

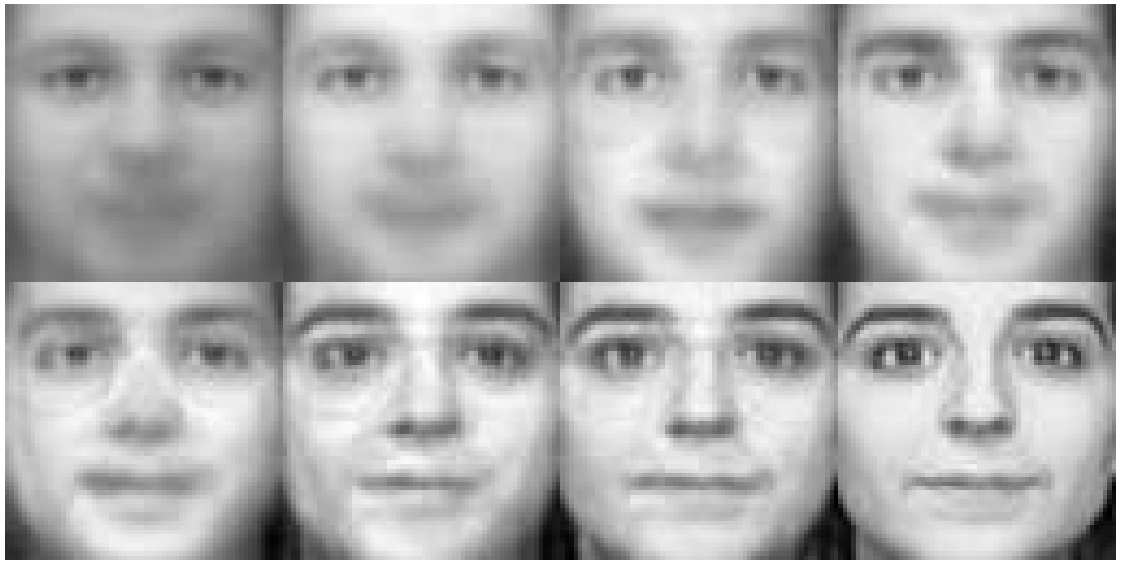
Observation index $i = 19$.

19th face. from top left to bottom right: (mean, 1, 5, 10) & (20, 50, 100, 400) PCs

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



- 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.
- 50 is still much smaller than 4096!
- Subjective and heuristic decision on "how many components to use"
- Reconstruction by PCA most useful and meaningful when
  - each datum is visually represented (rather than being just numbers).
  - for example: images, functions, shapes.

Handwritten Digits

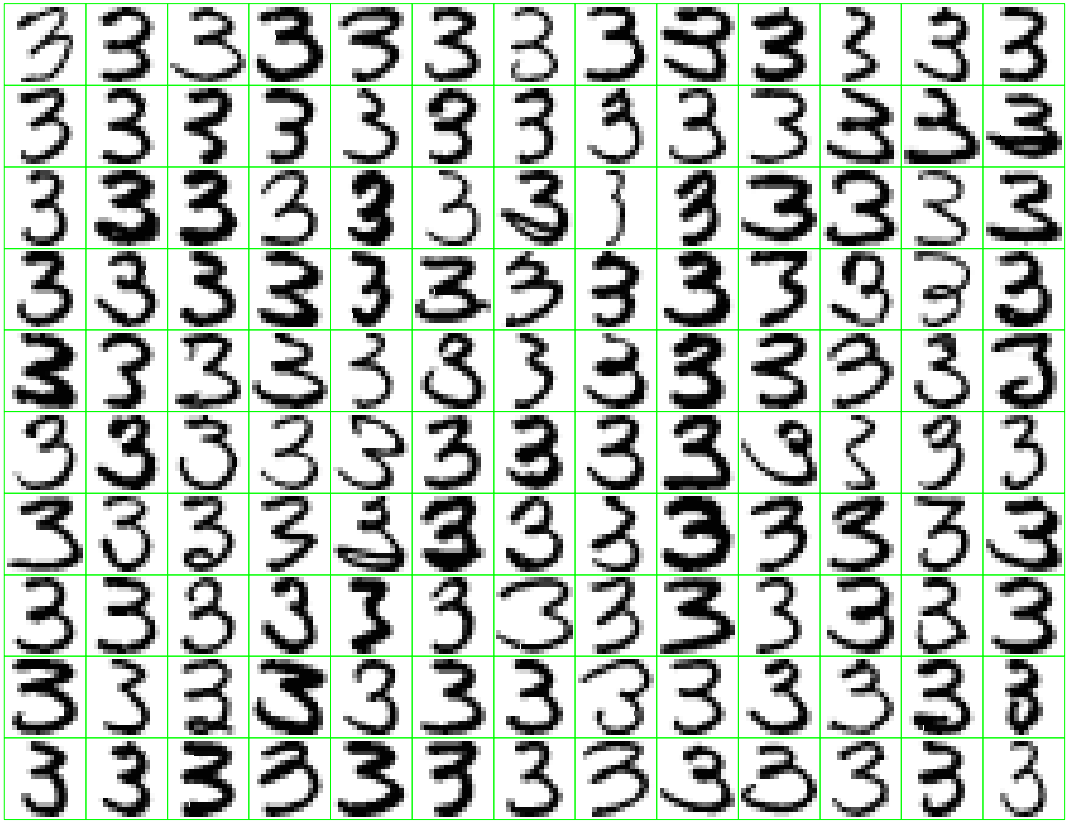


FIGURE 14.22. *A sample of* 130 *handwritten 3's shows a variety of writing styles.*

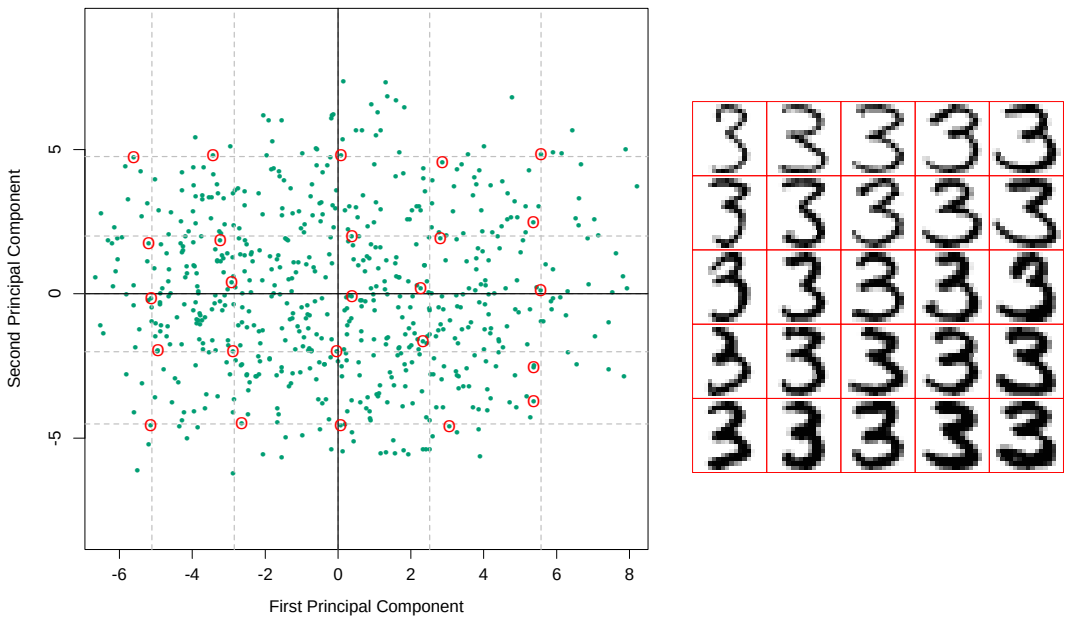


FIGURE 14.23. *(Left panel:) the first two principal components of the handwritten threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. (Right panel:) The images corresponding to the circled points. These show the nature of the first two principal components.*

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$$= \quad + \lambda_1 \cdot \quad + \lambda_2 \cdot \quad .$$

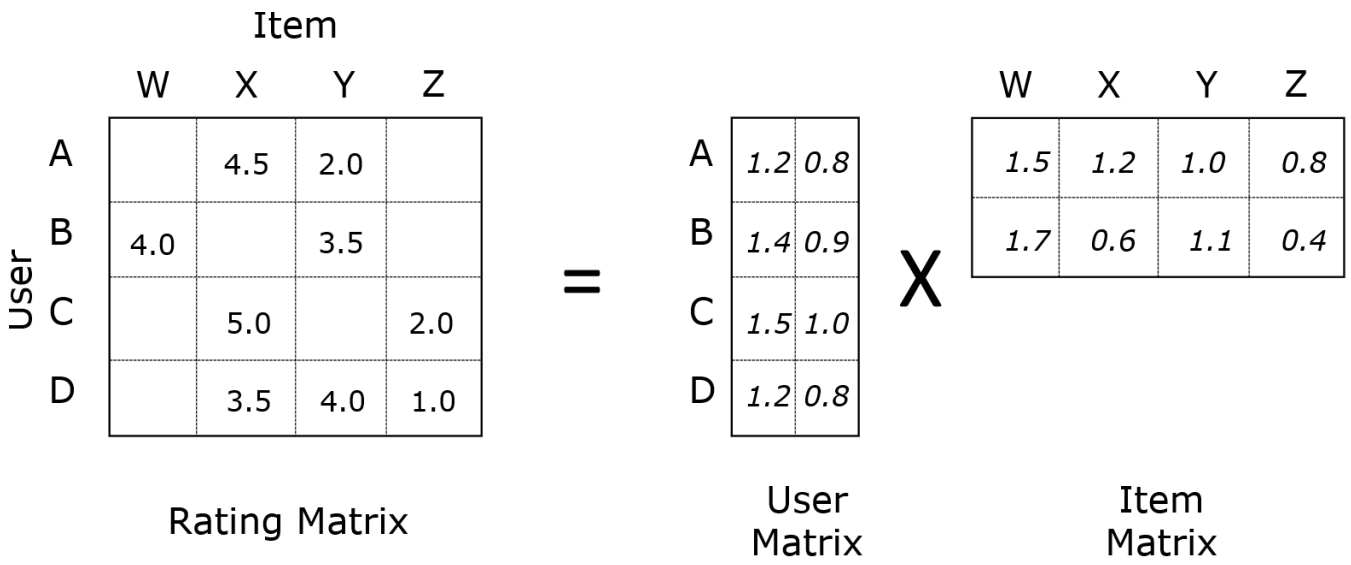
PCA as a mean of dimension reduction

- 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 $\mathbf{Z}_{d\times n}$.
- However,
  1. Unsupervised learning (no information on Y).
  2. Hard to interpret. Each PC (new variable) is a linear combination of p variables.
  3. Eigen-decomposition/SVD are problematic when $p \gg n$.

The next section would be . . . . . .

Matrix Completion

The matrix-completion problem has attracted a lot of attention, largely as a result of the celebrated Netflix Prize competition.



Foundation of collaborative filtering and recommendation system.

Candès and Tao (2009), Mazumder et al. (2010):

$$\min_{M} \frac{1}{2}\|(X - M)_{\Omega}\|_F^2 + \lambda\|M\|_*$$

where Ω is the set of available entries and $\|M\|_*$ is the nuclear norm which is the sum of the singular values of M

Rennie and Srebro (2005):

$$\min_{A,B} \frac{1}{2}\|(X - AB^T)_{\Omega}\|_F^2 + \lambda(\|A\|_F^2 + \|B\|_F^2)$$

where A and B have r columns.

Sparse PCA

Goal: PC directions should be sparse (many zero loadings)
Why? Better interpretation
Shen and Huang (2006): suppose rank =1

$$\underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}} \|\mathbf{X} - \boldsymbol{u}\boldsymbol{v}^T\|_F^2 + \lambda\|\boldsymbol{u}\|_1$$

subject to $\|\boldsymbol{v}\|_2 = 1$
Zou, Hastie and Tibshirani (2006):

$$\underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\theta}\boldsymbol{u}^T\boldsymbol{x}_i\|_2^2 + \lambda\|\boldsymbol{u}\|_2 + \lambda_1\|\boldsymbol{u}\|_1$$

subject to $\|\boldsymbol{\theta}\|_2 = 1$

Functional PCA

- Functional extension of PCA.
- FPCA. Suppose we observe functions $X_1(\cdot), X_2(\cdot), \cdots, X_n(\cdot)$. We want to find an orthonormal basis $\phi_1(\cdot), \cdots, \phi_K(\cdot)$ such that
$$\sum_{i=1}^{n} \|X_i - \sum_{k=1}^{K} \langle X_i, \phi_k \rangle \phi_k\|^2$$
is minimized.
- Once such a basis is found, we can replace each curve X_i by $\sum_{k=1}^{K} \langle X_i, \phi_k \rangle \phi_k$ as a good approximation.
- This means instead of working with infinitely dimensional curves X_i, we can work with K-dimensional vectors $(\langle X_i, \phi_1 \rangle, \cdots, \langle X_i, \phi_K \rangle)^\intercal$.

See Ramsay, J. and Silverman, B. (1997). Functional Data Analysis, Springer, New York.

PCA for Functional Data In Practice

- Each $X(\cdot)$ is observed at p times and stored as a p-dimensional vector
- n curves are organized as a $n \times p$ data matrix.
- Apply the regular PCA
- The p dimensional PC direction vector is converted to the eigenfunction $\phi_j(\cdot)$
- The PC scores are $\langle X_i, \phi_k \rangle$

Kernel PCA

- Goal: re-express PCA using inner products.
- Recall $\tilde{\mathbf{X}} = \mathbf{UDV}'$.
- Here $\mathbf{U}$ are the loadings and $\mathbf{Z} = \mathbf{DV}'$ is the PC scores.
- Let $\mathbf{K} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}} \in \mathbb{R}^{n\times n}$. Then $\mathbf{K} = \mathbf{VD}^2\mathbf{V}'$
- Conclusion: PCA = eigen decomposition of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = (\mathbb{I} - \mathbf{M})^T\mathbf{X}^T\mathbf{X}(\mathbb{I} - \mathbf{M})$
- Kernel PCA: eigen decomposition of $(\mathbb{I} - \mathbf{M})^T\mathbf{K}(\mathbb{I} - \mathbf{M})$, where $\mathbf{K}$ is the kernel matrix.

Note that there is no loading matrix (what is the PC direction vector in this case anyway?)

Supervised Dimension Reduction

- Partial Least Squares:
  - Best dimension reduction of cross-covariance between $\mathbf{X}$ and Y such that factors are orthogonal to $\mathbf{X}$.
- Canonical Correlations Analysis:
  - Best dimension reduction of cross-covariance between $\mathbf{X}$ and $\mathbf{Y}$ such that bi-projection is orthogonal to $\mathbf{X}$ or $\mathbf{Y}$.
- Linear Discriminant Analysis (classification):
  - Best dimension reduction of between class covariance matrix relative to within-class covariance.