

# Dimension reduction 1: Principal component analysis

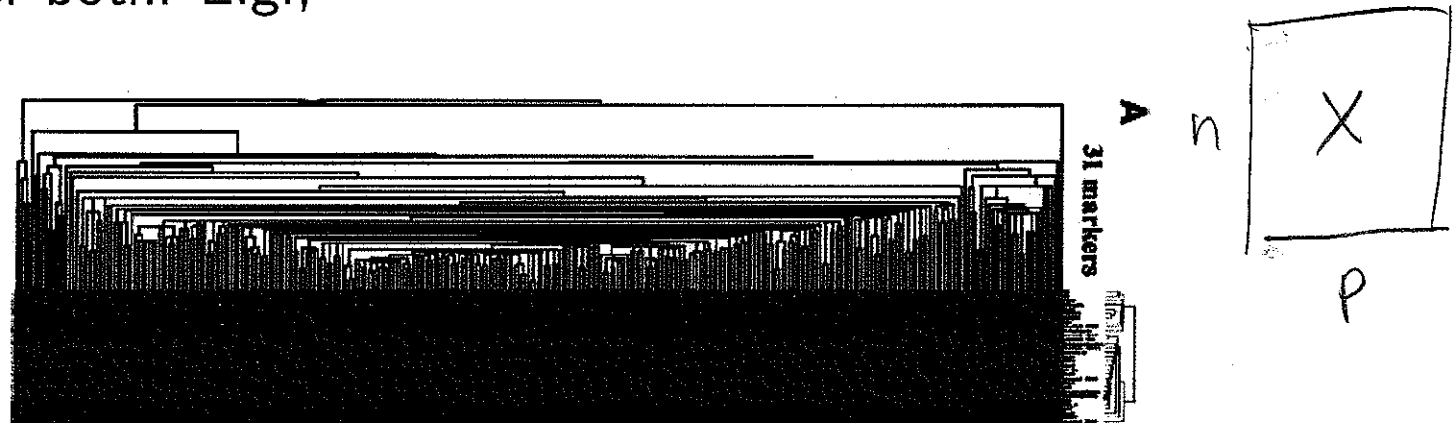
Ryan Tibshirani  
Data Mining: 36-462/36-662

February 5 2013

*Optional reading: ISL 10.2, ESL 14.5*

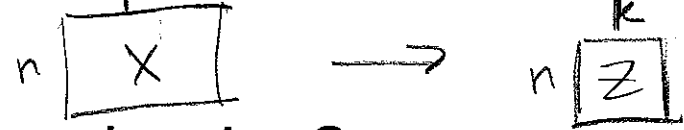
# Clustering as dimension reduction

We've thought about clustering observations, given features. But in many situations, we can actually cluster the observations or the features or both. E.g.,



(From Makretsov et al. (2004), "Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma")

If we cluster the features using  $K$ -means or hierarchical clustering, then we could replace the features by cluster centers. This would reduce the dimension of our feature space



## What is dimension reduction?

Dimension reduction: the task of transforming our data set to one with less features. A new feature can be one of the old features, or it can be a some linear or nonlinear combination of old features. We want this transformation to preserve the main structure that is present in the feature space

This is a broader goal than that of clustering. It is often the first step in an analysis, to be followed by, e.g., visualization, clustering, regression, classification

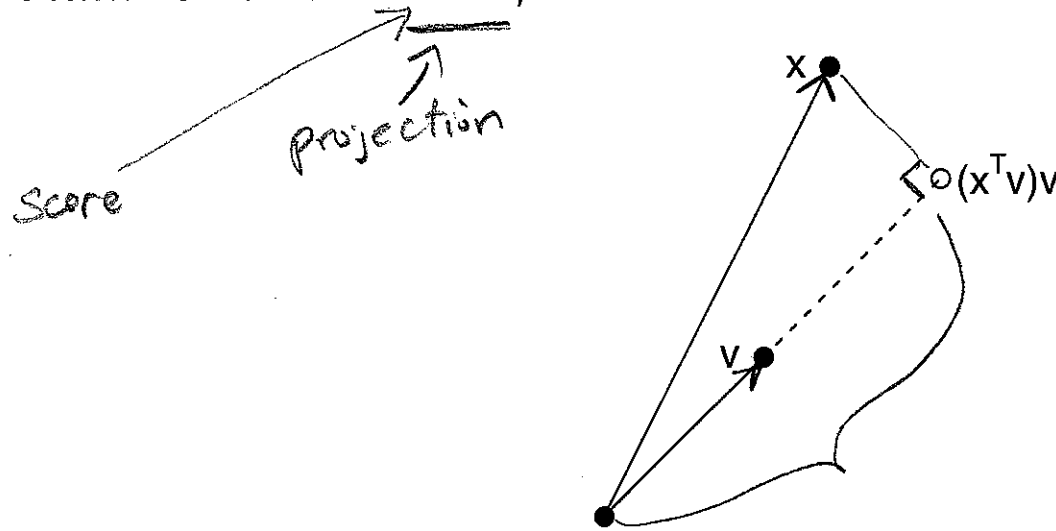
straight lines in  $\mathbb{R}^d$   $x_1, \dots, x_n \in \mathbb{R}^d$

We're going to start with linear dimension reduction. This means: looking for straight lines in the feature space along which the data exhibit an interesting trend

Specifically, we're going to interpret "interesting" to mean high variance

## Review: projections onto unit vectors

A vector  $v \in \mathbb{R}^p$  with  $\|v\|_2^2 = v^T v = 1$  is said to have unit norm. The projection of  $x \in \mathbb{R}^p$  onto (the direction of)  $v$  is  $(x^T v)v$ . Think of this as  $c \cdot v$ , with a coefficient or "score" of  $c = x^T v$



Consider a matrix  $X \in \mathbb{R}^{n \times p}$ . and consider projecting each row

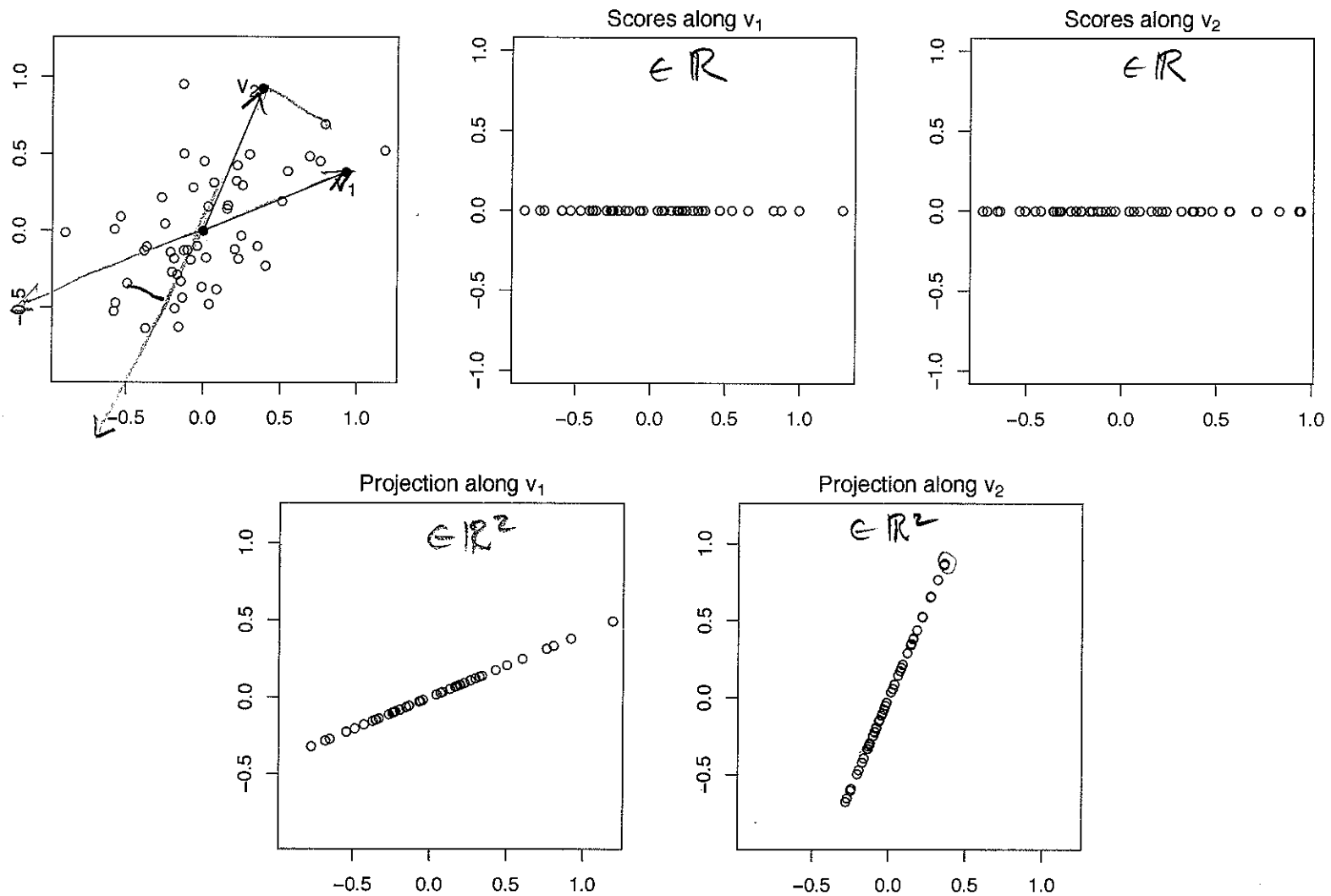
$x_i \in \mathbb{R}^p$  onto  $\underbrace{v}_{\text{unit vector}}$ . The entries of  $\underline{Xv} = \begin{pmatrix} x_1^T v \\ x_2^T v \\ \vdots \\ x_n^T v \end{pmatrix} \in \mathbb{R}^n$  are the

scores, and the rows of  $\underline{Xvv^T} \in \mathbb{R}^{n \times p}$  are the projected vectors

$$\begin{pmatrix} x_1^T v & v v^T \\ x_2^T v & v v^T \\ \vdots & \vdots \end{pmatrix}$$

# Example: projections onto unit vectors

Example:  $X \in \mathbb{R}^{50 \times 2}$ ,  $v_1, v_2 \in \mathbb{R}^2$



## Review: projections onto orthonormal vectors

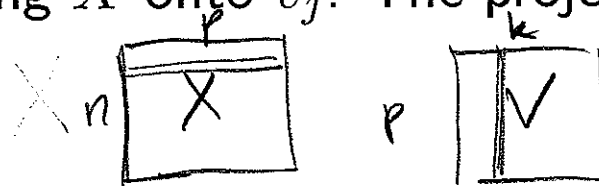
Vectors  $v_1, v_2 \in \mathbb{R}^p$  are orthogonal if  $v_1^T v_2 = 0$ , and  $v_1, \dots, v_k \in \mathbb{R}^p$  are orthogonal if  $v_i^T v_j = 0$  for any  $i, j$ . Vectors  $v_1, \dots, v_k \in \mathbb{R}^p$  are orthonormal if they are orthogonal and each  $v_j$  has unit norm

The projection of  $x \in \mathbb{R}^p$  onto (the space spanned by) orthonormal vectors  $v_1, \dots, v_k \in \mathbb{R}^p$  is  $\sum_{j=1}^k (x^T v_j) v_j$ . The score along the  $j$ th direction is  $x^T v_j$ .   
*linear combos of*  
 $\sum c_j v_j$

Write the collection  $v_1, \dots, v_k \in \mathbb{R}^p$  as a matrix  $V \in \mathbb{R}^{p \times k}$ , where each  $v_j$  is a column. Consider a data matrix  $X \in \mathbb{R}^{n \times p}$ , we want to project rows of  $X$  onto columns of  $V$ . The scores are given by

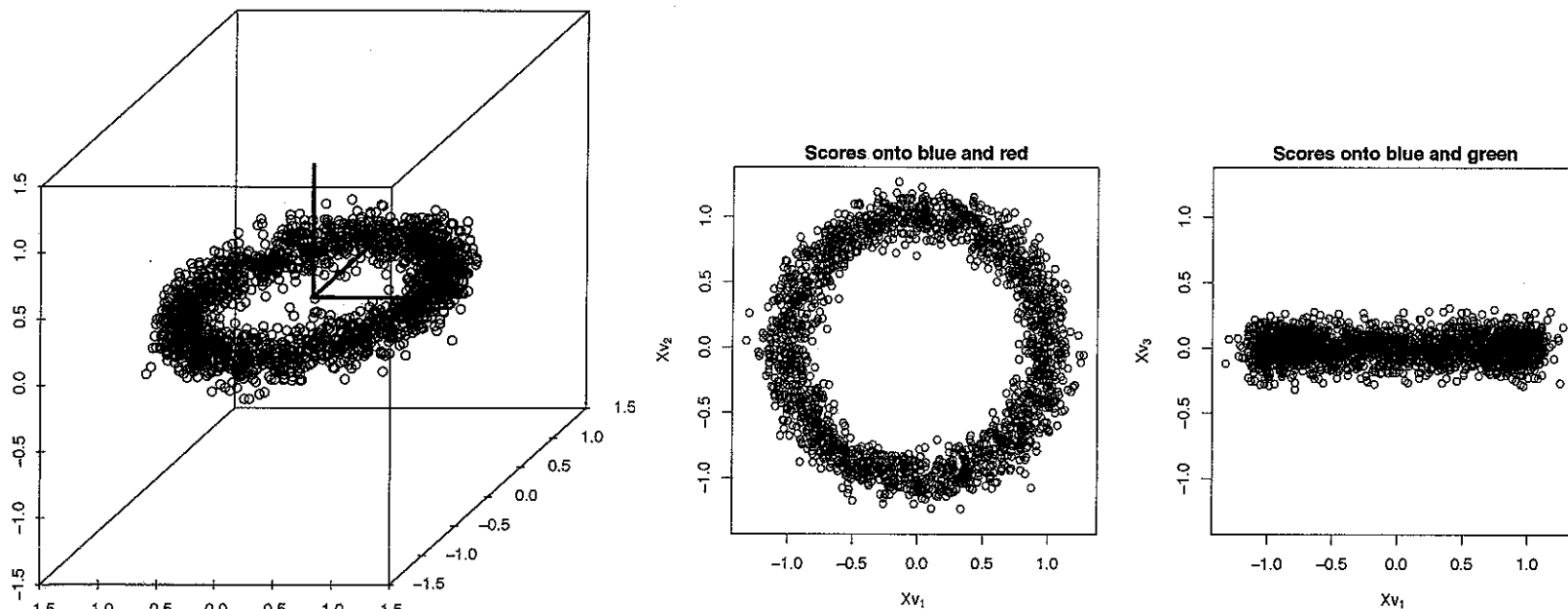
$XV \in \mathbb{R}^{n \times k}$ , with  $j$ th column  $Xv_j = \begin{pmatrix} x_1^T v_j \\ x_2^T v_j \\ \vdots \\ x_n^T v_j \end{pmatrix} \in \mathbb{R}^n$ , which  
*scores*

contains the scores from projecting  $X$  onto  $v_j$ . The projections are the rows of  $XVV^T \in \mathbb{R}^{n \times p}$ .  
*projected vectors*



## Example: projections onto orthonormal vectors

Example:  $X \in \mathbb{R}^{2000 \times 3}$ , and  $v_1, v_2, v_3 \in \mathbb{R}^3$  are the unit vectors parallel to the coordinate axes



Not all linear projections are equal! What makes a good one?

Exactly what PCA is looking for...  
high variance  
(total sample variance)

## Review: sample statistics (in vector notation)

Given a vector  $x \in \mathbb{R}^n$  of  $n$  observations  $\frac{1}{n} \sum_{i=1}^n x_i$

Sample mean:  $\bar{x} = \frac{1}{n} x^T \mathbf{1} \in \mathbb{R}$ , where  $\mathbf{1} \in \mathbb{R}^n$  is the vector of 1s

Sample variance:  $\frac{1}{n} (x - \bar{x} \mathbf{1})^T (x - \bar{x} \mathbf{1}) \in \mathbb{R}$   
 $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

Given a matrix  $X \in \mathbb{R}^{n \times p}$ , of  $n$  observations by  $p$  features

Sample mean vector:  $\bar{X} = \frac{1}{n} X^T \mathbf{1} \in \mathbb{R}^p$   $\bar{X} = \left( \frac{1}{n} \sum_{i=1}^n x_{i1}, \frac{1}{n} \sum_{i=1}^n x_{i2}, \dots, \right)$

Sample covariance matrix:  $\frac{1}{n} (X - \mathbf{1} \bar{X}^T)^T (X - \mathbf{1} \bar{X}^T) \in \mathbb{R}^{p \times p}$

Total sample variance:  $\text{trace} \left( \frac{1}{n} (X - \mathbf{1} \bar{X}^T)^T (X - \mathbf{1} \bar{X}^T) \right) \in \mathbb{R}$

(where the trace is simply the sum of the diagonal entries, i.e., for  $A \in \mathbb{R}^{p \times p}$ ,  $\text{trace}(A) = \sum_{i=1}^p A_{ii}$ )



## Centering vectors and matrices

$$(x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$$

To center  $x \in \mathbb{R}^n$  means to replace it by  $\tilde{x} = x - \bar{x}\mathbf{1} \in \mathbb{R}^n$ . The new  $\tilde{x}$  has sample mean zero, but its sample variance is the same as before:  $\frac{1}{n}\tilde{x}^T\tilde{x} = \frac{1}{n}(x - \bar{x}\mathbf{1})^T(x - \bar{x}\mathbf{1})$   $\frac{1}{n}\tilde{x}^T\mathbf{1} = \frac{1}{n}(x^T\mathbf{1} - \bar{x}\mathbf{1}^T\mathbf{1}) = 0$

To center (or column-center)  $X \in \mathbb{R}^{n \times p}$  means to replace it by  $\tilde{X} = X - \mathbf{1}\bar{X}^T \in \mathbb{R}^{n \times p}$ . Each column of  $\tilde{X}$  now has sample mean zero, but the sample covariance of  $\tilde{X}$  is the same as before:  $\frac{1}{n}\tilde{X}^T\tilde{X} = \frac{1}{n}(X - \mathbf{1}\bar{X}^T)^T(X - \mathbf{1}\bar{X}^T)$

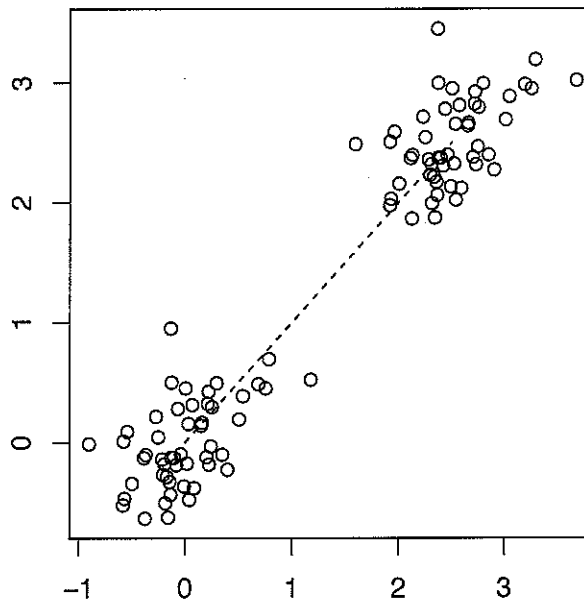
Assume that the columns of  $X \in \mathbb{R}^{n \times p}$  have been centered (drop the tilde notation). Then  $Xv \in \mathbb{R}^n$  has sample mean zero for any vector  $v \in \mathbb{R}^p$  (Homework 2), therefore the sample variance of  $Xv$  is  $\frac{1}{n}(Xv)^T(Xv) = \frac{1}{n}\|Xv\|_2^2$

(Centering makes the math cleaner!)

# Principal component analysis

Principal component analysis (PCA) is nearly as old as statistics itself. Because it has been widely studied, you will hear it being called different things in different fields

We are given a data matrix  $X \in \mathbb{R}^{n \times p}$ , meaning that we have  $n$  observations (row vectors) and  $p$  features (column vectors). We assume that the columns of  $X$  have been centered. (Is this going to change the structure that we're interested in?)



# First principal component direction and score

The first principal component direction of  $X$  is the unit vector  $\underline{v_1} \in \mathbb{R}^p$  that maximizes the sample variance of  $\underline{Xv_1} \in \mathbb{R}^n$  when compared to all other unit vectors

Score when project onto  $v_1$

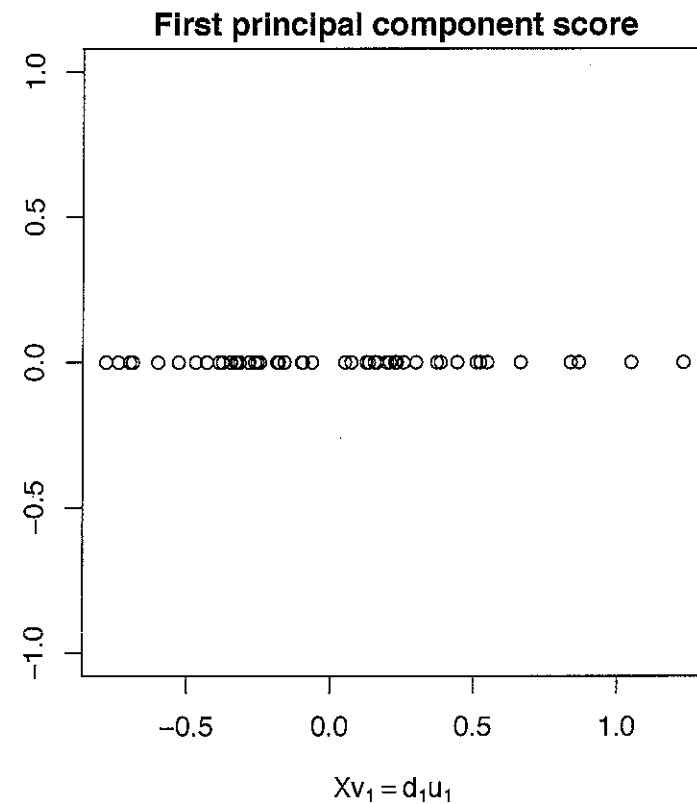
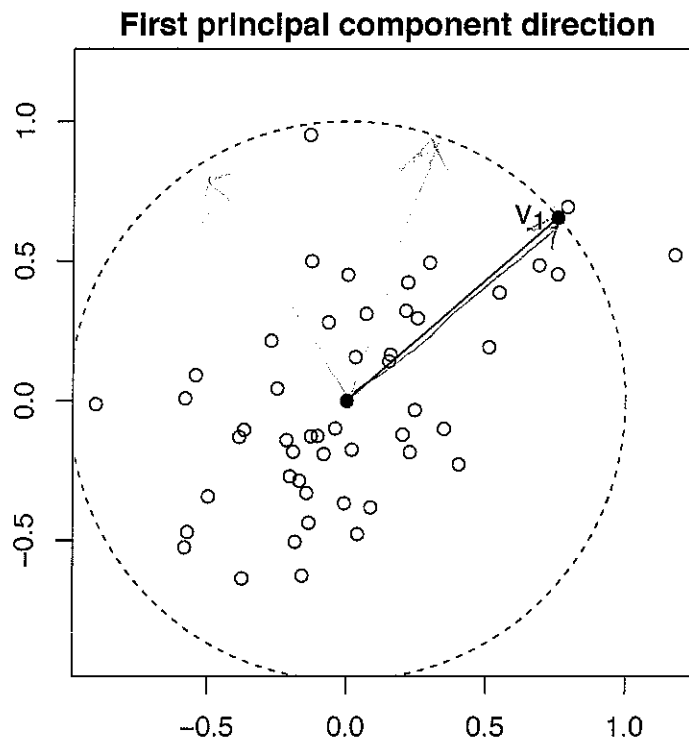
For any  $v \in \mathbb{R}^p$ , the vector  $Xv \in \mathbb{R}^n$  has sample mean zero and sample variance  $\frac{1}{n}(Xv)^T(Xv)$  (recall that we column centered  $X$ ). Hence the first principal component direction  $v_1 \in \mathbb{R}^p$  is

$$\text{direction} \rightarrow v_1 = \operatorname{argmax}_{\|v\|_2=1} \underbrace{(Xv)^T(Xv)}_{\text{sample variance of } Xv}$$

The vector  $\underline{Xv_1} \in \mathbb{R}^n$  is called the first principal component score of  $X$ , and  $u_1 = (Xv_1)/d_1 \in \mathbb{R}^n$  is the normalized first principal component score. Here  $d_1 = \sqrt{(Xv_1)^T(Xv_1)}$ , and  $\underline{d_1^2/n}$  is the amount of variance explained by  $v_1$

# Example: first principal component direction and score

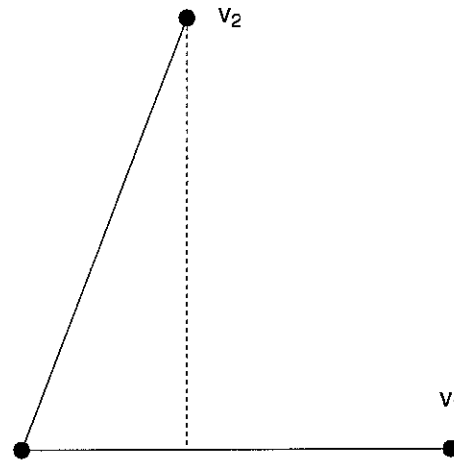
Same example data as earlier:  $X \in \mathbb{R}^{50 \times 2}$



## Beyond the first direction and score

What happens next? The idea is to successively find orthogonal directions of the highest variance.

Why orthogonal? Because we've already explained the variance in  $X$  along  $v_1$ , and now we want to look at variance in a different direction. Any direction not orthogonal to  $v_1$  would necessarily have some overlap with  $v_1$ , i.e., it would create some redundancy in explaining the variance in  $X$ .



(Plus, it makes the math easier!)

## Second principal component direction and score

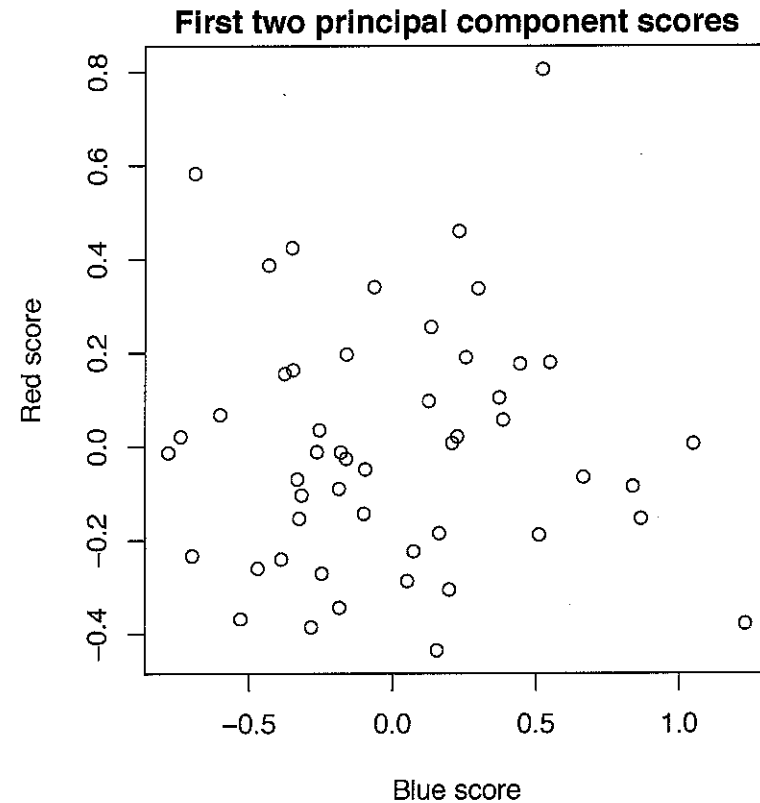
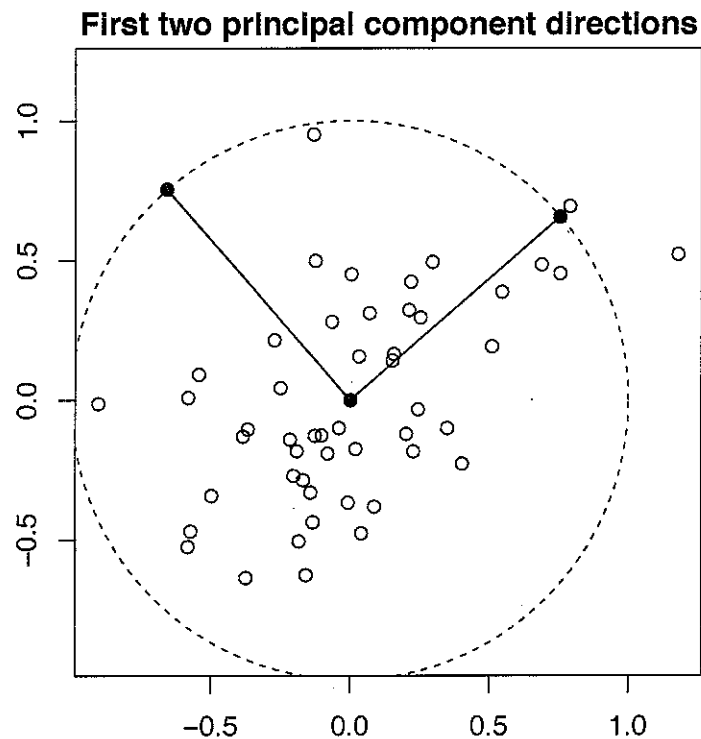
Given the first principal component direction  $v_1 \in \mathbb{R}^p$ , we define the second principal component direction  $v_2 \in \mathbb{R}^p$  to be the unit vector, with  $v_2^T v_1 = 0$ , that makes  $Xv_2 \in \mathbb{R}^n$  have maximal sample variance over all unit vectors orthogonal to  $v_1$ . This is

$$\text{direction} \nearrow v_2 = \underset{\substack{\|v\|_2=1 \\ v^T v_1=0}}{\operatorname{argmax}} \underbrace{(Xv)^T (Xv)}_{\text{sample variance}}$$

The vector  $Xv_2 \in \mathbb{R}^n$  is called the second principal component score of  $X$ , and  $u_2 = (Xv_2)/d_2 \in \mathbb{R}^n$  is the normalized second principal component score. Here  $d_2 = \sqrt{(Xv_2)^T (Xv_2)}$ , and  $d_2^2/n$  is the amount of variance explained by  $v_2$

# Example: second principal component direction and score

Same example data as earlier:  $X \in \mathbb{R}^{50 \times 2}$



## Further principal component directions and scores

Given the  $k - 1$  principal component directions  $v_1, \dots, v_{k-1} \in \mathbb{R}^p$  (note that these are orthonormal), we define the  $k$ th principal component direction  $v_k \in \mathbb{R}^p$  to be

$$\underline{v_k} = \underset{\substack{\|v\|_2=1 \\ v^T v_j = 0, j=1, \dots, k-1}}{\operatorname{argmax}} \quad \underbrace{(Xv)^T (Xv)}_{\substack{\uparrow \text{sample variance}}}$$

The vector  $Xv_k \in \mathbb{R}^n$  is called the  $k$ th principal component score of  $X$ , and  $u_k = (Xv_k)/d_k \in \mathbb{R}^n$  is the normalized  $k$ th principal component score. Here  $d_k = \sqrt{(Xv_k)^T (Xv_k)}$ , and  $d_k^2/n$  is the amount of variance explained by  $v_k$

$$\frac{d_1^2}{n} \geq \frac{d_2^2}{n} \geq \dots \geq \frac{d_k^2}{n}$$



# Properties and representations

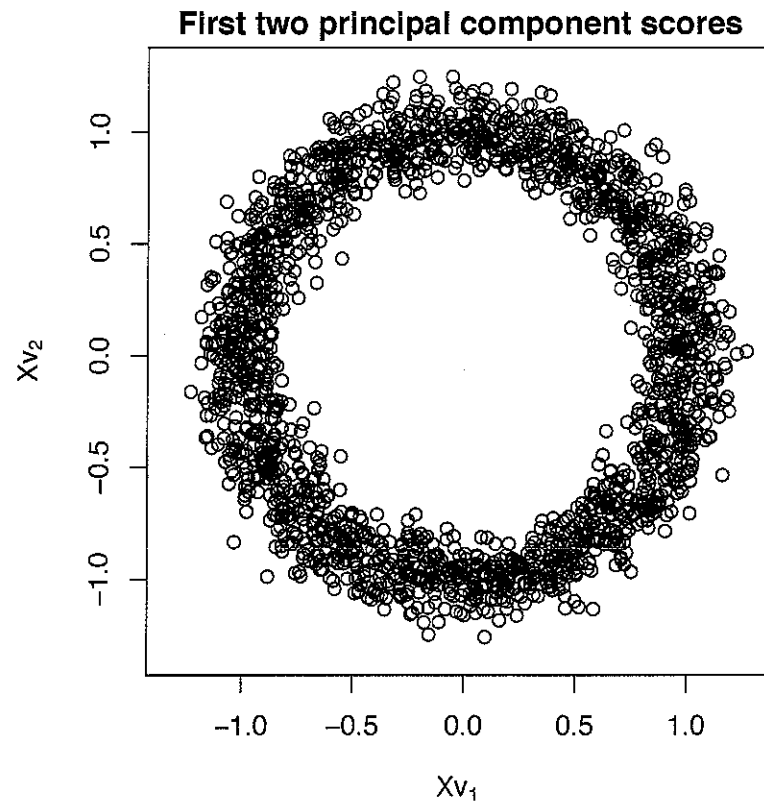
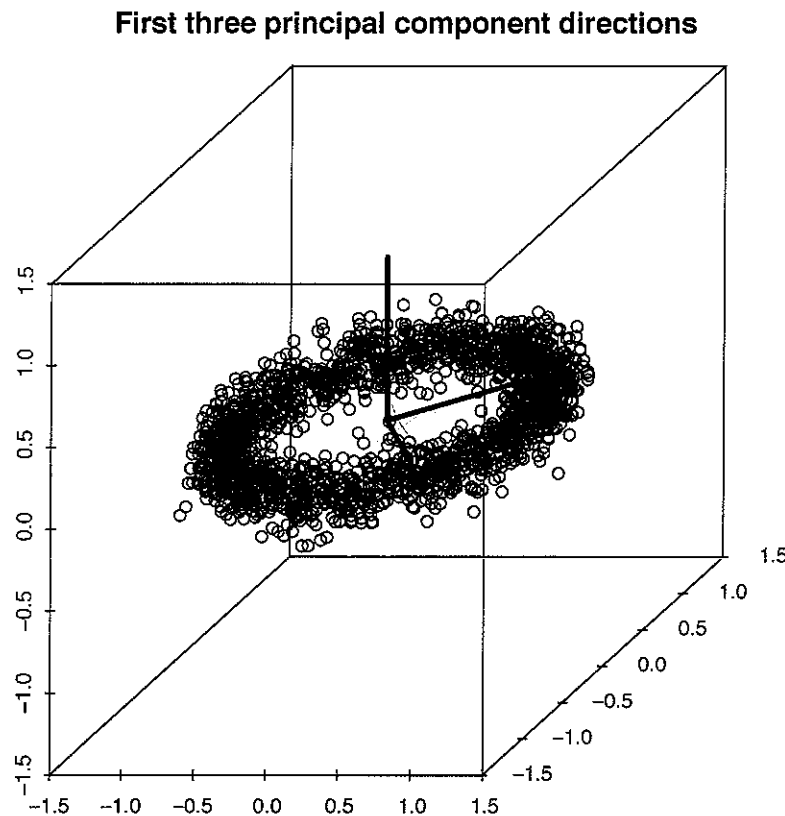
$$UDV^T$$

- ▶ For the  $k$ th principal component direction  $v_k \in \mathbb{R}^p$  and score  $u_k \in \mathbb{R}^n$ , the entries of  $Xv_k = d_k u_k$  are the scores from projecting  $X$  onto  $v_k$ , and the rows of  $Xv_k v_k^T = d_k u_k v_k^T$  are the projected vectors
- ▶ The directions  $v_k$  and normalized scores  $u_k$  are only unique up to sign flips
- ▶ How many principal component directions/scores are there? There are  $p$ , because if  $\underline{v_1, \dots, v_p} \in \mathbb{R}^p$  are orthonormal, then they are linearly independent<sup>1</sup>
- ▶ Concise representation: let the columns of  $V \in \mathbb{R}^{p \times p}$  be the directions. Scores: columns of  $XV \in \mathbb{R}^{n \times p}$ . Projections onto  $V_k$  (first  $k$  columns of  $V$ ): rows of  $\underline{XV_k V_k^T} \in \mathbb{R}^{n \times p}$   
 $\rightarrow [Xv_1 \ Xv_2 \ \dots \ Xv_p] \in \mathbb{R}^{n \times k}$ 
 $XV_k \in \mathbb{R}^{n \times k}$

<sup>1</sup>To be precise, here we are assuming that  $p \leq n$  and  $\text{rank}(X) = p$ . In general, there are exactly  $r = \text{rank}(X)$  principal component directions

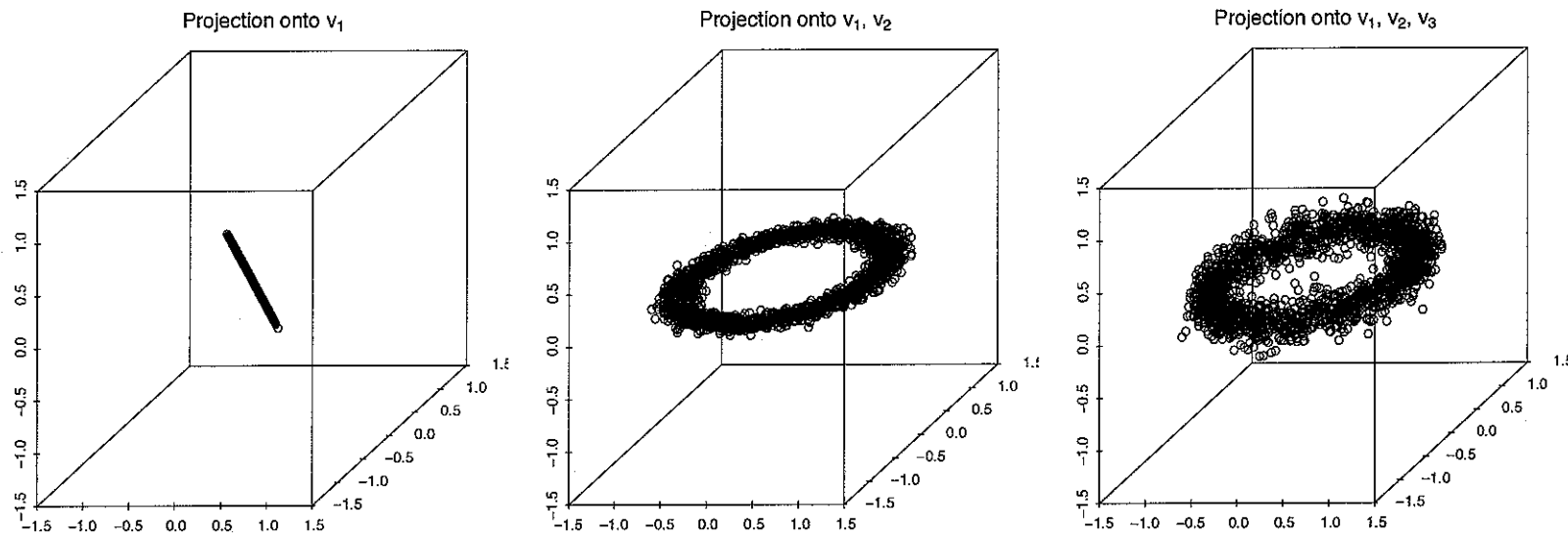
## Example: principal component analysis in $\mathbb{R}^3$

Example:  $X \in \mathbb{R}^{2000 \times 3}$ . Shown are the three principal component directions  $v_1, v_2, v_3 \in \mathbb{R}^3$ , and the scores from projecting onto the first two directions



# Example: projecting onto principal component directions

Same example:  $X \in \mathbb{R}^{2000 \times 3}$ ,  $v_1, v_2, \dots, v_3 \in \mathbb{R}^3$ . What happens if replace  $X$  by its projection onto  $v_1$ ? Onto  $v_1, v_2$ ? Onto  $v_1, v_2, v_3$ ?



The third plot looks exactly the same as the original data. Is this a coincidence? No! (Why?)

$$X V_k V_k^T$$

## Proportion of variance explained

Recall that we said:  $d_k^2/n$  is the amount of variance explained by the  $k$ th principal component direction  $v_k$

Two facts (Homework 2):

- ▶ The total sample variance of  $X$  is  $\frac{1}{n} \sum_{j=1}^p d_j^2$
- ▶ The total sample variance of  $XV_kV_k^T$  is  $\frac{1}{n} \sum_{j=1}^k d_j^2$  (amount of variance explained by  $v_1 \dots v_k$ )

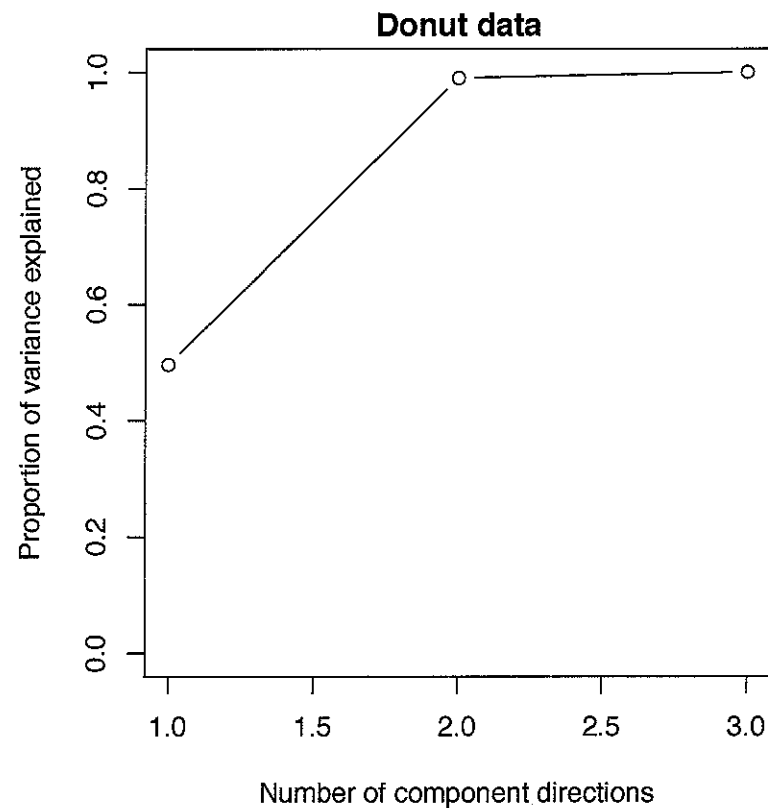
Hence the proportion of variance explained by the first  $k$  principal component directions  $v_1, \dots, v_k$  is

$$\frac{\sum_{j=1}^k d_j^2}{\sum_{j=1}^p d_j^2}$$

If this is high for a small value of  $k$ , then it means that the main structure in  $X$  can be explained by a small number of directions

## Example: proportion of variance explained

Example: proportion of variance explained as a function of  $k$ , for the donut data



# Principal component analysis in R


The function `princomp` in the base package computes directions and scores via an eigendecomposition of  $X^T X$ . E.g.,

```
pc = princomp(x)
dirs = pc$loadings # directions
scrs = pc$scores   # scores
```

The function `prcomp` in the base package computes directions and scores via a singular value decomposition of  $X$ . E.g.,

E.g.,

```
pc = prcomp(x)
dirs = pc$rotation
scrs = pc$x
```



## Recap: principal component analysis

We reviewed basic projective geometry, and sample statistics in vector/matrix notation

We defined the principal component directions  $v_1, \dots, v_p \in \mathbb{R}^p$  of a centered matrix  $X \in \mathbb{R}^{n \times p}$ , as successively orthogonal unit vectors that maximize the sample variance

We also defined the principal component scores  $Xv_1 = d_1u_1, \dots, Xv_p = d_pu_p \in \mathbb{R}^n$ , and the amounts of variance explained by each direction  $d_1^2/n, \dots, d_p^2/n$

The proportion of variance explained is a nice way to quantify how much structure is being captured as  $k$  varies

# Next time: more principal component analysis

How do we actually compute principal component directions and scores? What can we do with them?

