Lecture 3: Principle Component Analysis (PCA)

dimensional; linear; very important
put on the resume
find a lower dimensional solution for the data set

Reading: Section 14.5

GU4241/GR5241 Statistical Machine Learning

X is an nxp matrix this is an unsupervised learning
there are two ways to find best low dimensional approximation
1. singular value decomposition（考）
x = UDV"t (u1,d1 are the score variable, d1^2 is the variance of the first col, d2^2 var sed col)
2. eigen decomposition（考）
1/n*X`tX =====X`t X= VD^2
<Ui,Uj> = {1, i = j; 0,i=/=j}

Linxi Liu

January 22, 2017

nxp nxp pxp pxp
X = U D V"t

Xv1 = u1 d
XV"t = UD
Xv2 = u2d2

The principle of the good low dim approximation:
then we need to get the sum of the distance, find the direction which minimize the distance
average squared lence

var(Xv2) = d1"tu1d1 = d1^2

for PCA, always centralized data, make the mean as the center

di^2 is the largest one , decreasing order

Eigenvalues

We consider a square matrix $A \in \mathbb{R}^{m \times m}$.

Definition

A vector $\xi \in \mathbb{R}^m$ is called an **eigenvector** of A if the direction of ξ does not change under application of A. In other words, if there is a scalar λ such that

$$A\xi = \lambda\xi \ .$$

λ is called an **eigenvalue** of A for the eigenvector ξ.

Properties in general

- In general, eigenvalues are complex numbers $\lambda \in \mathbb{C}$.
- The class of matrices with the nicest eigen-structure are symmetric matrices, for which all eigenvectors are mutually orthogonal.

Eigenstructure of symmetric matrices

If a matrix is symmetric:

- There are rank(A) distinct eigendirections.
- The eigenvectors are pair-wise orthogonal.
- If rank$(A) = m$, there is an ONB of $\mathbb{R}^m$ consisting of eigenvectors of A.

Definiteness

type	if ...
positive definite	all eigenvalues > 0
positive semi-definite	all eigenvalues ≥ 0
negative semi-definite	all eigenvalues ≤ 0
negative definite	all eigenvalues < 0
indefinite	none of the above

Orthonormal Bases

Recall: ONB

A basis $\{v_1, \ldots, v_m\}$ of $\mathbb{R}^m$ is called an **orthonormal basis** if

$$\langle v_i, v_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

In other words, the v_i are pairwise orthogonal and each of length 1.

Orthogonal matrices

A matrix is orthogonal precisely if its rows form an ONB. Any two ONBs can be transformed into each other by an orthogonal matrix.

Transforming between ONBs

If $\mathcal{V} = \{v_1, \ldots, v_m\}$ and $\mathcal{W} = \{w_1, \ldots, w_m\}$ are ONBs, there is an orthogonal matrix O such that

$$A_{[\mathcal{V}]} = O A_{[\mathcal{W}]} O^{-1}$$

for any matrix A. By $A_{[\mathcal{V}]}$, we denote the representation of A in $\mathcal{V}$.

Eigenvector ONB

Setting

- Suppose A symmetric, $\xi_1, \ldots, \xi_m$ are eigenvectors and form an ONB.
- $\lambda_1, \ldots, \lambda_m$ are the corresponding eigenvalues.

How does A act on a vector $v \in \mathbb{R}^m$?

1. Represent v in basis $\xi_1, \ldots, \xi_m$:

$$v = \sum_{j=1}^{m} v_j^{\mathsf{A}} \xi_j \qquad \text{where } v_j^{\mathsf{A}} \in \mathbb{R}$$

2. Multiply by A: Eigenvector definition (recall: $A\xi_j = \lambda\xi_j$) yields

$$Av \quad = \quad A\Big(\sum_{j=1}^{m} v_j^{\mathsf{A}} \xi_j\Big) \quad = \quad \sum_{j=1}^{m} v_j^{\mathsf{A}} A\xi_j \quad = \quad \sum_{j=1}^{m} v_j^{\mathsf{A}} \lambda_j \xi_j$$

Conclusion

A symmetric matrix acts by scaling the directions ξ_j.

Illustration

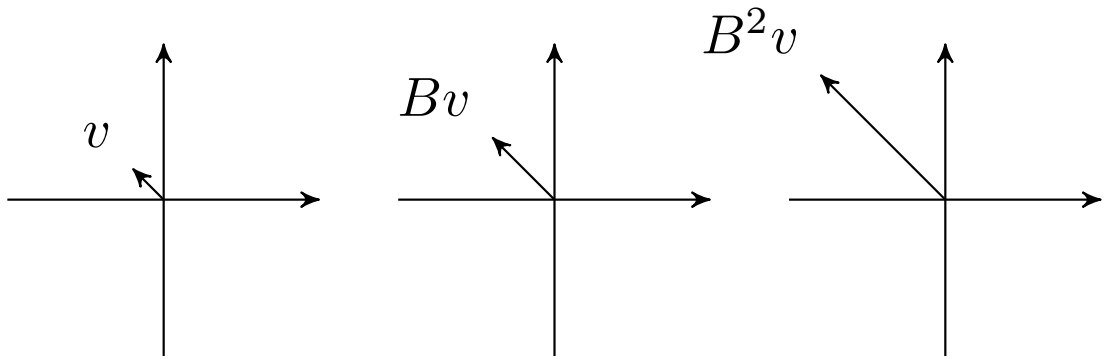
Setting

We *repeatedly* apply a symmetric matrix B to some vector $v \in \mathbb{R}^m$, i.e. we compute

$$Bv, \quad B(Bv) = B^2v, \quad B(B(Bv))) = B^3v, \quad \dots$$

How does v change?

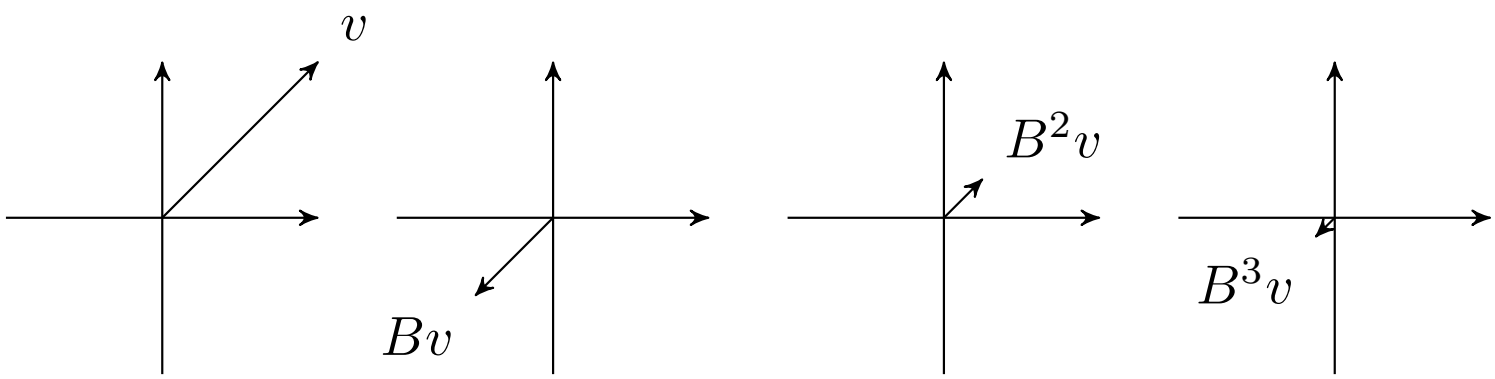
Example 1: v is an eigenvector with eigenvalue 2



The direction of v does not change, but its length doubles with each application of B.

Illustration

Example 2: v is an eigenvector with eigenvalue $-\frac{1}{2}$



For an arbitrary vector v

$$B^n v = \sum_{j=1}^{m} v_j \mathsf{B} \lambda_j^n \xi_j$$

- The weight λ_j^n grows most rapidly for eigenvalue with largest absolute value.
- Consequence:

  The direction of $B^n v$ converges to the direction of the eigenvector with largest eigenvalue as n grows large.

Quadratic Forms

In applications, symmetric matrices often occur in quadratic forms.

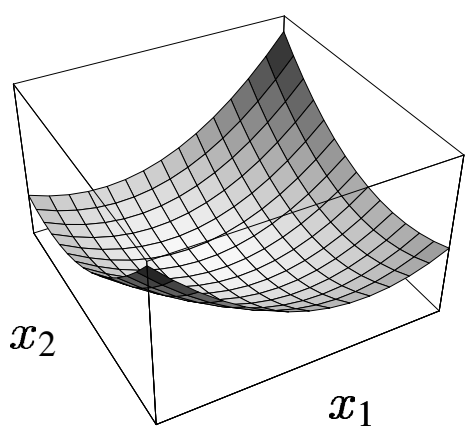
Definition

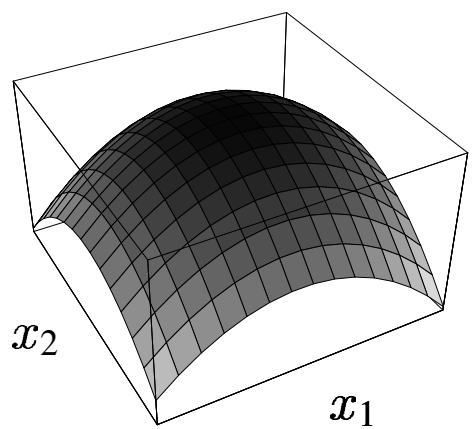
The **quadratic form** defined by a matrix A is the function

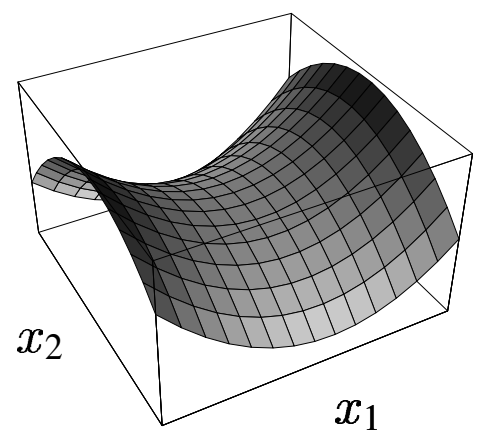
$$\begin{aligned} q_A : \mathbb{R}^m &\to \mathbb{R} \\ x &\mapsto \langle x, Ax \rangle \end{aligned}$$

Intuition

A quadratic form is the m-dimensional analogue of a quadratic function ax^2, with a vector substituted for the scalar x and the matrix A substituted for the scalar $a \in \mathbb{R}$.

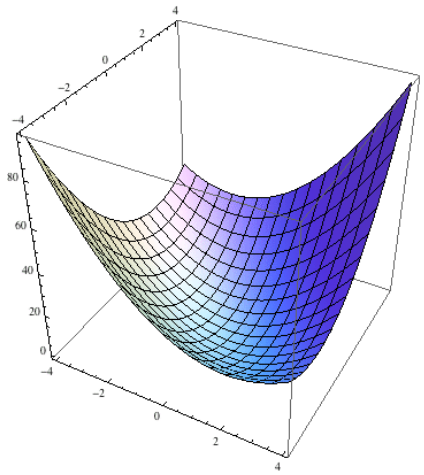


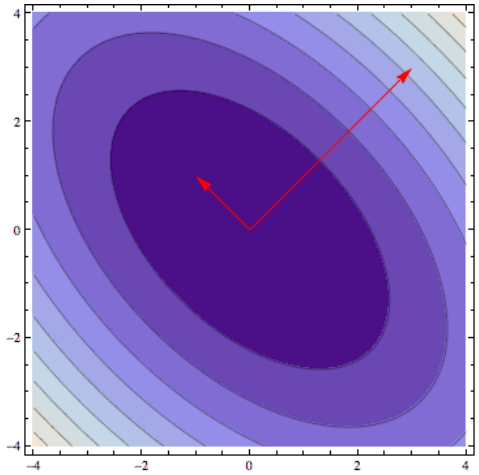




Quadratic Forms

Here is the quadratic form for the matrix $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$:

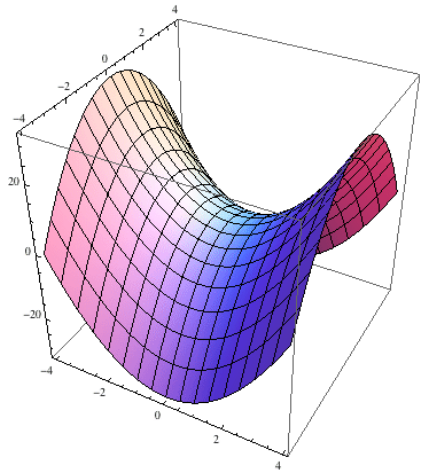


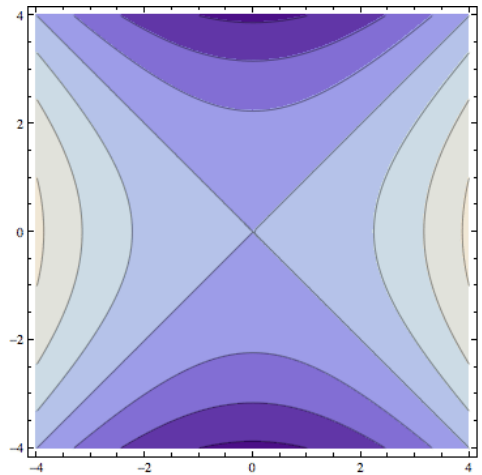


- Left: The function value q_A is graphed on the vertical axis.
- Right: Each line corresponds to a constant function value of q_A.
  Dark color = small values.
- The red lines are eigenvector directions of A. Their lengths represent the (absolute) values of the eigenvalues.
- In this case, both eigenvalues are positive. If all eigenvalues are positive, the contours are ellipses. So:
  positive definite matrices $\leftrightarrow$ elliptic quadratic forms

Quadratic Forms

In this plot, the eigenvectors are axis-parallel, and one eigenvalue is negative:





The matrix here is $A = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$.

Intuition

- If we change the sign of one of the eigenvalue, the quadratic function along the corresponding eigen-axis flips.
- There is a point which is a minimum of the function along one axis direction, and a maximum along the other. Such a point is called a *saddle point*.

Application: Covariance Matrix

Recall: Covariance

The covariance of two random variables X_1, X_2 is

$$\mathrm{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] \, .$$

If $X_1 = X_2$, the covariance is the variance: $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$.

Covariance matrix

If $X = (X_1, \ldots, X_m)$ is a random vector with values in $\mathbb{R}^m$, the matrix of all covariances

$$\mathrm{Cov}[X] := (\mathrm{Cov}[X_i, X_j])_{i,j} = \begin{pmatrix} \mathrm{Cov}[X_1, X_1] & \cdots & \mathrm{Cov}[X_1, X_m] \\ \vdots & & \vdots \\ \mathrm{Cov}[X_m, X_1] & \cdots & \mathrm{Cov}[X_m, X_m] \end{pmatrix}$$

is called the **covariance matrix** of X.

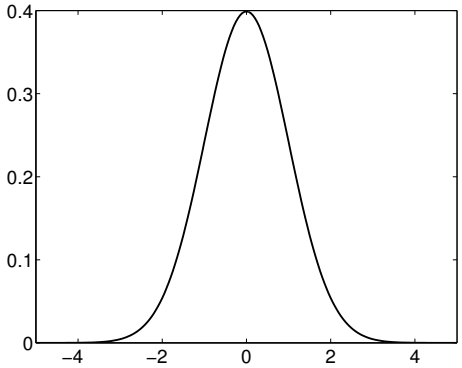
Notation

It is customary to denote the covariance matrix $\mathrm{Cov}[X]$ by Σ.

Gaussian Distribution

Gaussian density in one dimension

$$p(x;\mu,\sigma) := \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$$

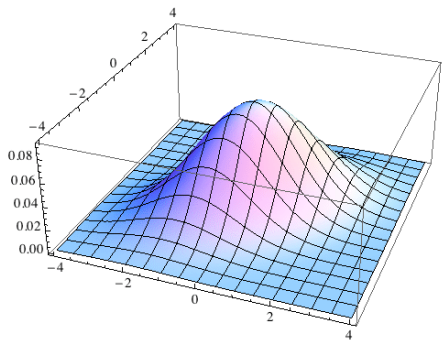


- μ = expected value of x, σ^2 = variance, σ = standard deviation
- The quotient $\frac{x-\mu}{\sigma}$ measures deviation of x from its expected value in units of σ (i.e. σ defines the length scale)

Gaussian density in m dimensions

The quadratric function

$$-\frac{(x-\mu)^2}{2\sigma^2} \quad = \quad -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$

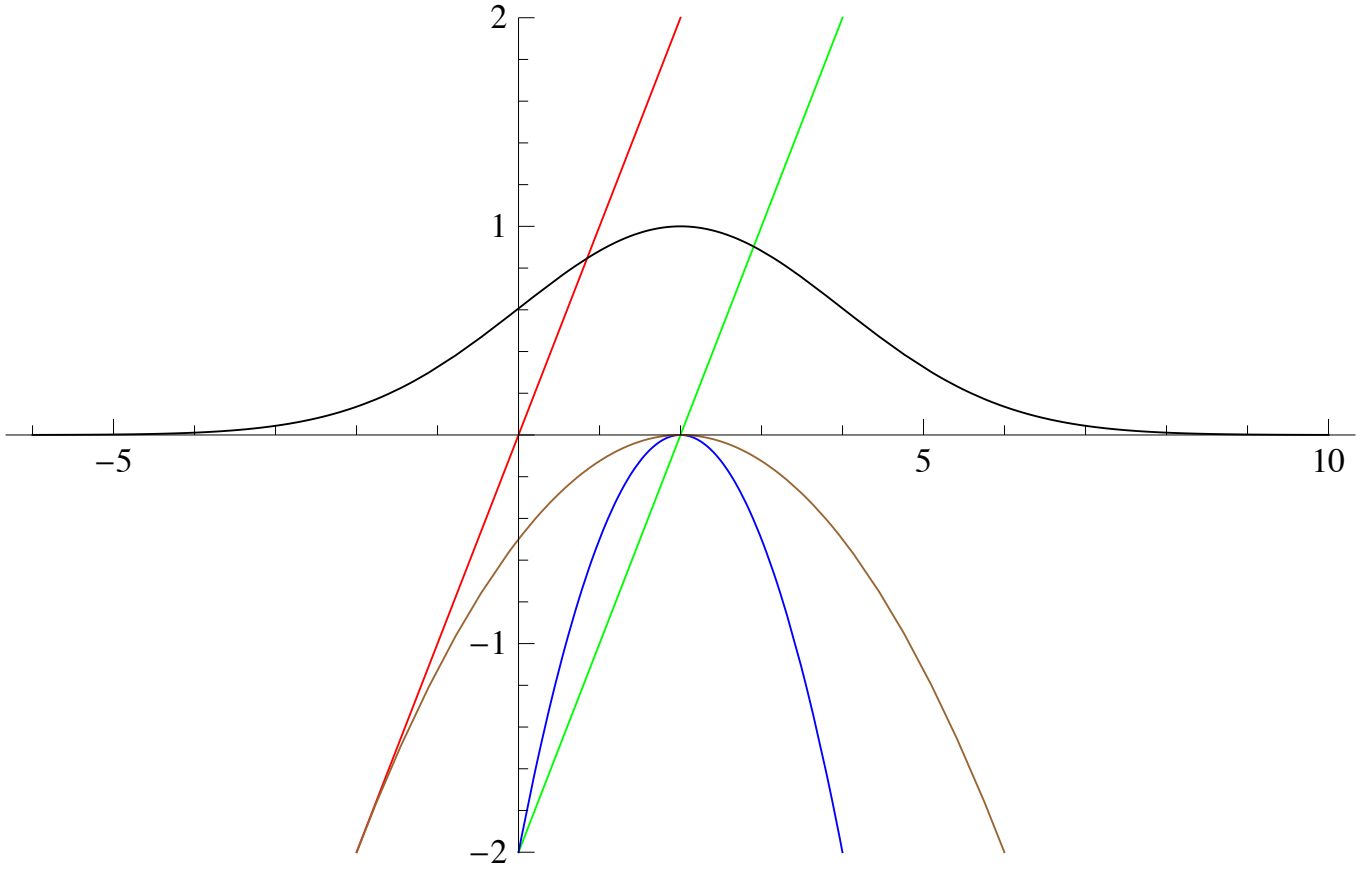


is replaced by a quadratic form:

$$p(\mathbf{x};\boldsymbol{\mu},\Sigma) := \frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\Big(-\frac{1}{2}\left\langle(\mathbf{x}-\boldsymbol{\mu}),\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\rangle\Big)$$

Components of a 1D Gaussian

$\mu = 2,\ \sigma = 2$



- Red: $x \mapsto x$
- Green: $x \mapsto x - \mu$
- Blue: $x \mapsto -\frac{1}{2}(x-\mu)^2$
- Brown: $x \mapsto -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$
- Black: $x \mapsto \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

Geometry of Gaussians

Covariance matrix of a Gaussian

If a random vector $X \in \mathbb{R}^m$ has Gaussian distribution with density $p(\mathbf{x}; \mu, \Sigma)$, its covariance matrix is $\mathrm{Cov}[X] = \Sigma$. In other words, a Gaussian is parameterized by its covariance.

Observation

Since $\mathrm{Cov}[X_i, X_j] = \mathrm{Cov}[X_j, X_i]$, the covariance matrix is symmetric.

What is the eigenstructure of Σ?

- We know: Σ symmetric $\Rightarrow$ there is an eigenvector ONB
- Call the eigenvectors in this ONB $\xi_1, \ldots, \xi_m$ and their eigenvalues $\lambda_1, \ldots, \lambda_m$
- We can rotate the coordinate system to $\xi_1, \ldots, \xi_m$. In the new coordinate system, Σ has the form

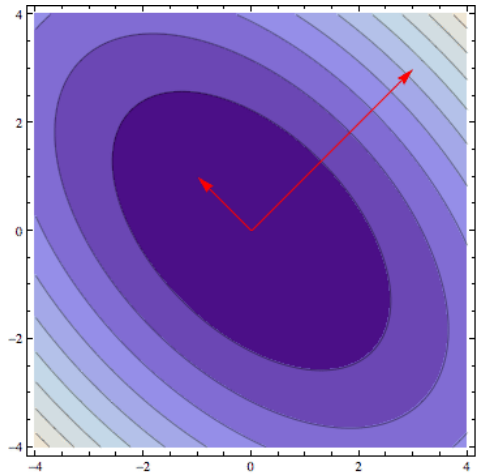
$$\Sigma_{[\xi_1,\ldots,\xi_n]} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$$

Example

Quadratic form

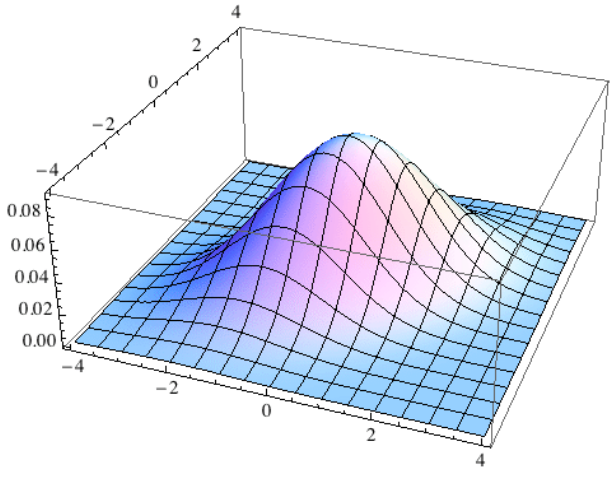
$$\langle \mathbf{x}, \Sigma \mathbf{x} \rangle \quad \text{with} \quad \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

The eigenvectors are $(1, 1)$ and $(-1, 1)$ with eigenvalues 3 and 1.

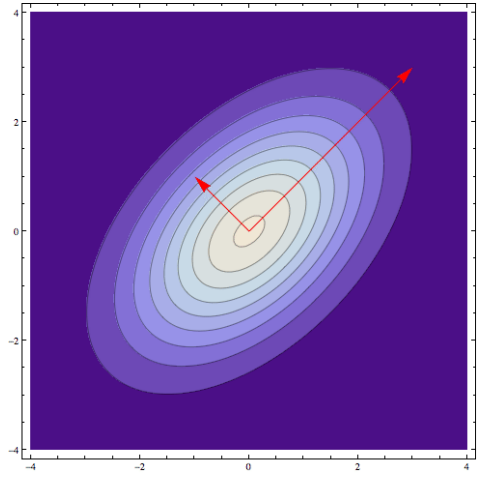


Gaussian density

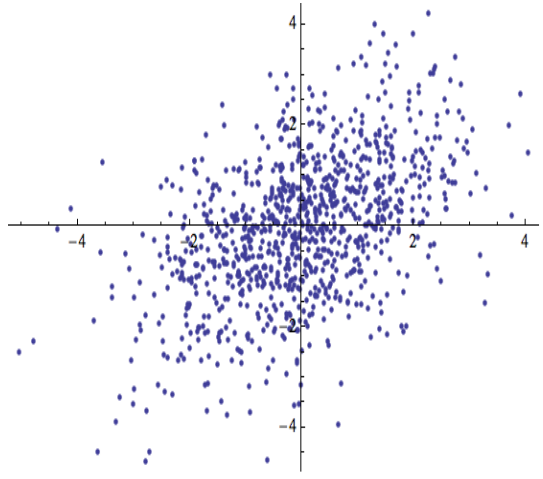
$p(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (0, 0)$.



Density graph



Density contour



1000 sample points

Interpretation

The ξ_i as random variables

Write $e_1, \ldots, e_m$ for the ONB of axis vectors. We can represent each ξ_i as

$$\xi_i = \sum_{j=1}^{m} \alpha_{ij} e_j$$

Then $O = (\alpha_{ij})$ is the orthogonal transformation matrix between the two bases.
We can represent random vector $X \in \mathbb{R}^m$ sampled from the Gaussian in the eigen-ONB as

$$X_{[\xi_1,\ldots,\xi_m]} = (X'_1, \ldots, X'_m) \qquad \text{with} \qquad X'_i = \sum_{j=1}^{m} \alpha_{ij} X_j$$

Since the X_j are random variables (and the α_{ij} are fixed), each X'_i is a scalar random variable.

Interpretation

Meaning of the random variables ξ_i

For any Gaussian $p(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$, we can

1. shift the origin of the coordinate system into $\boldsymbol{\mu}$
2. rotate the coordinate system to the eigen-ONB of Σ.

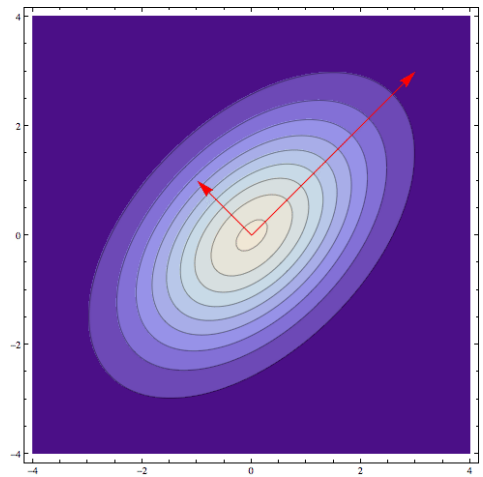
In this new coordinate system, the Gaussian has covariance matrix

$$\Sigma_{[\xi_1,\ldots,\xi_m]} = \mathsf{diag}(\lambda_1, \ldots, \lambda_m)$$

where λ_i are the eigenvalues of Σ.

Gaussian in the new coordinates

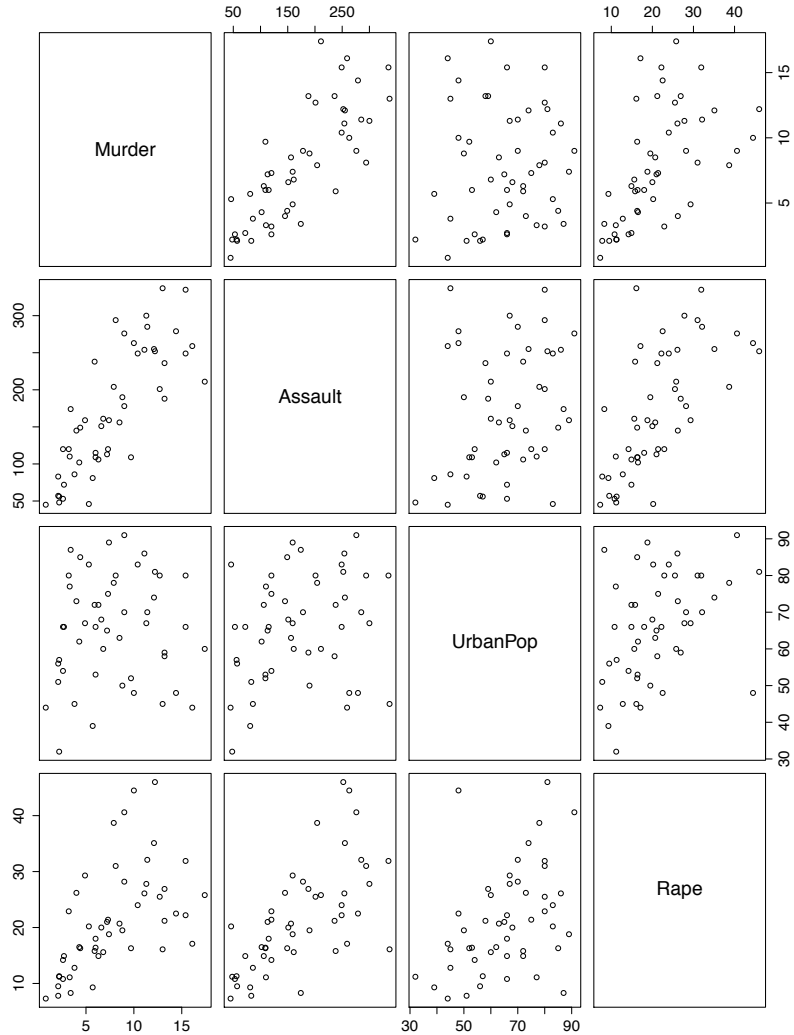
A Gaussian vector $X_{[\xi_1,\ldots,\xi_m]}$ represented in the new coordinates consists of m *independent* 1D Gaussian variables X'_i. Each X'_i has mean 0 and variance λ_i.



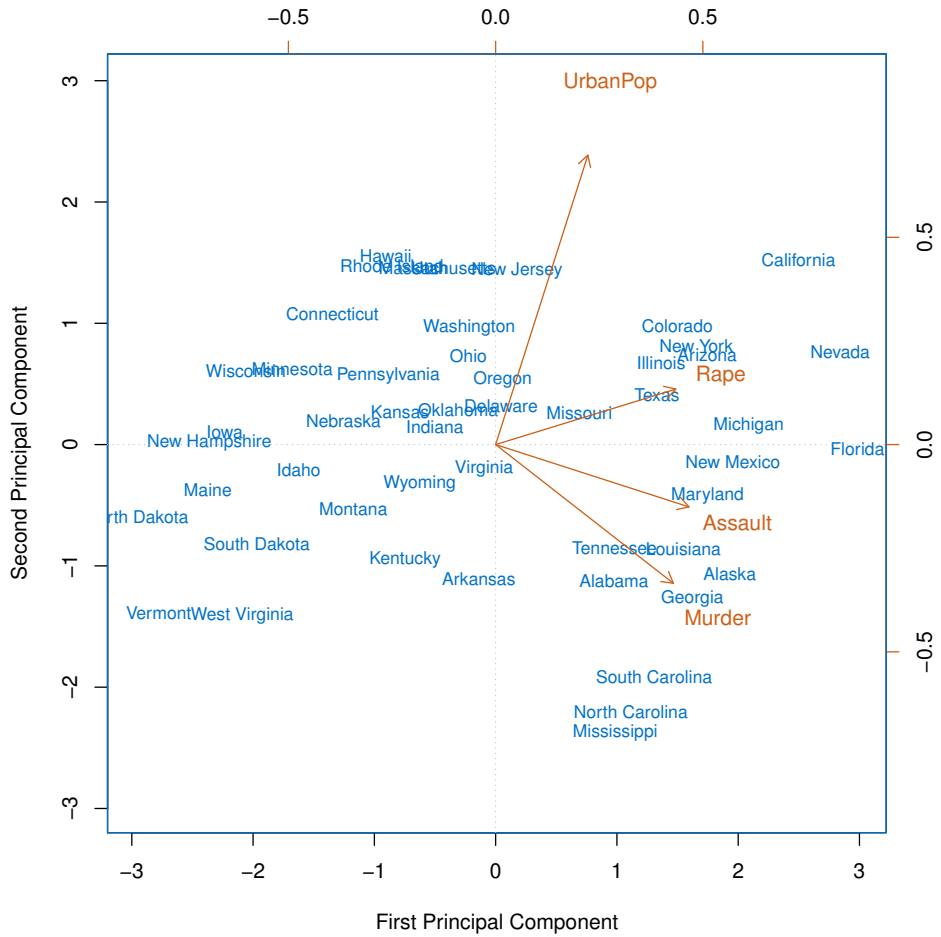
Principal Component Analysis

- This is the most popular unsupervised procedure ever.
- Invented by Karl Pearson (1901).
- Developed by Harold Hotelling (1933).
- **What does it do?** It provides a way to visualize high dimensional data, summarizing the most important information.

What is PCA good for?



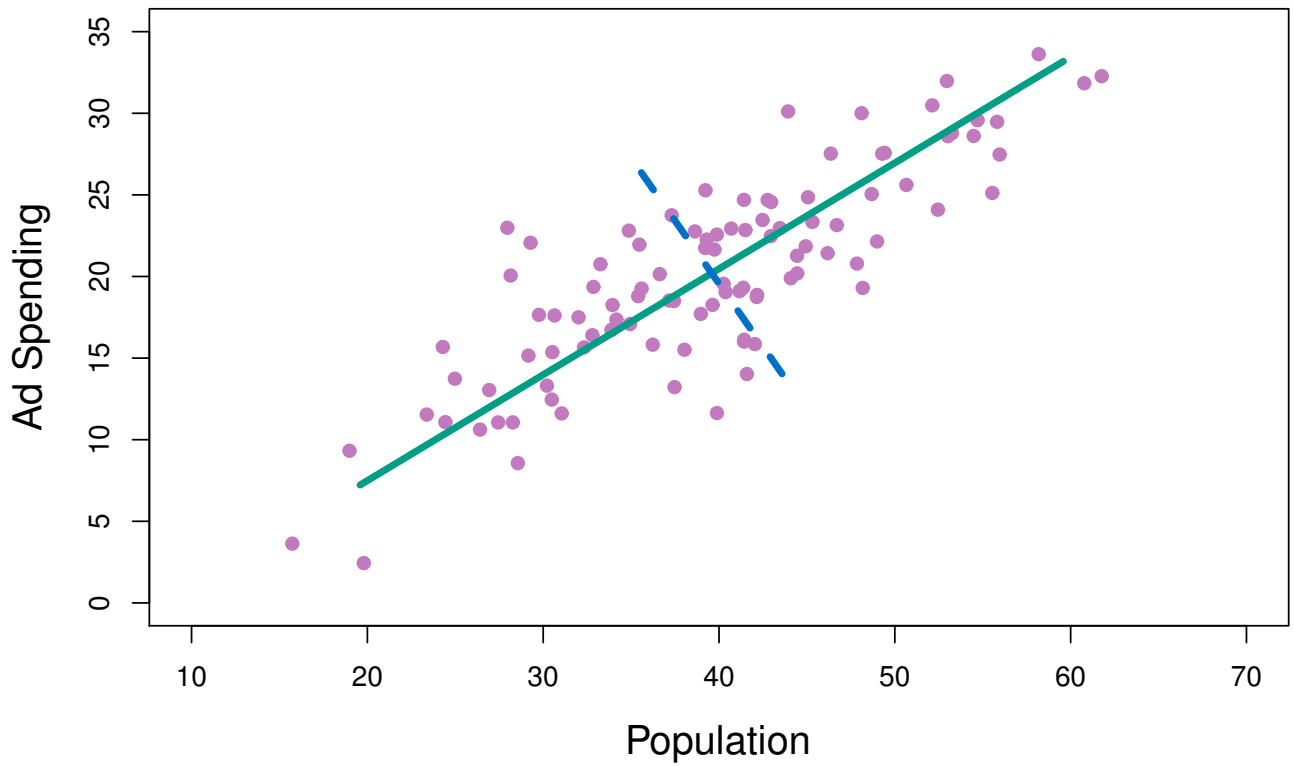
What is PCA good for?



ISL Figure 10.1

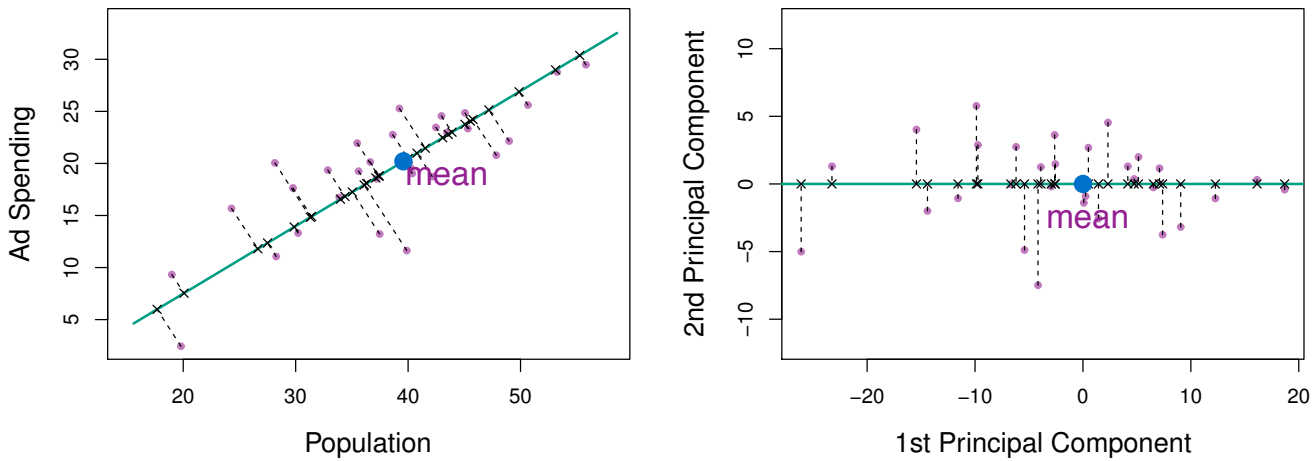
What is the first principal component?

It is the vector which passes the closest to a cloud of samples, in terms of squared Euclidean distance.



i.e. The green direction minimizes the average squared length of the dotted lines.

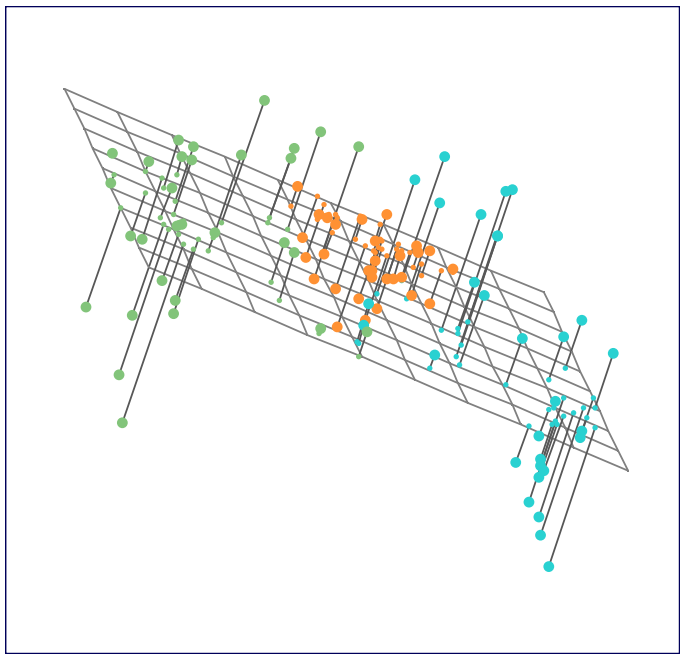
the distance to the mean is fixed.

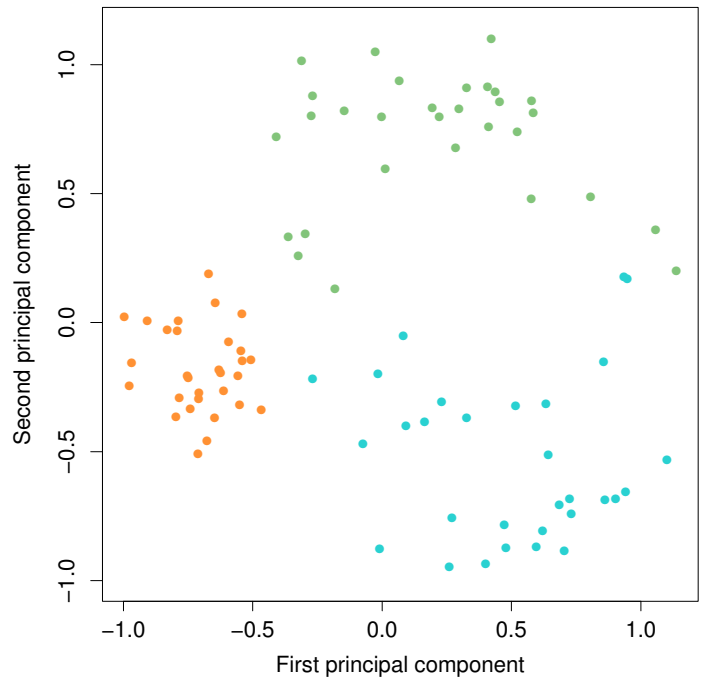


ISL Figure 6.15

What does this look like with 3 variables?

The first two principal components span a plane which is closest to the data.

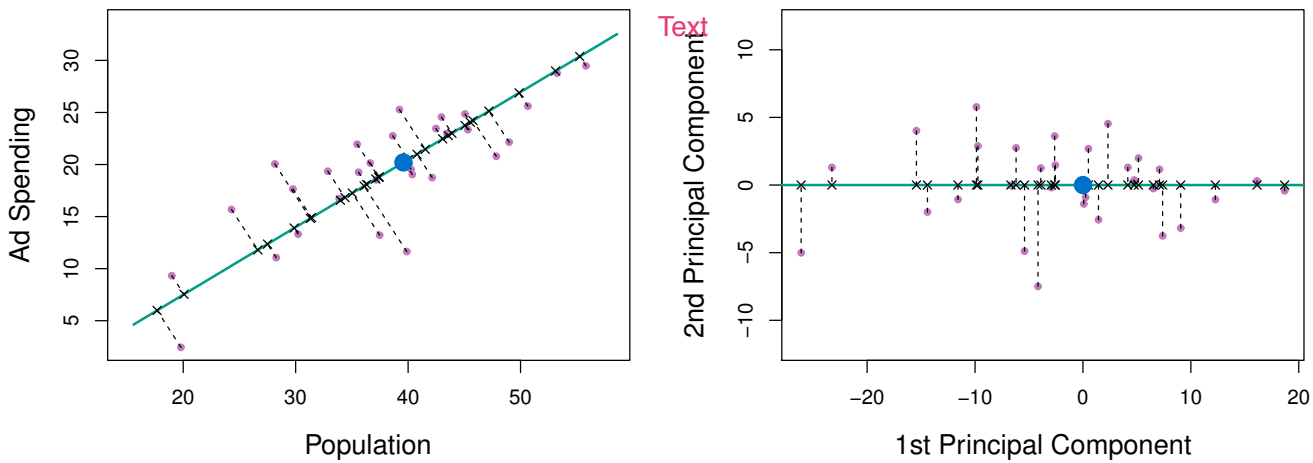




ISL Figure 10.2

A second interpretation

The projection onto the first principal component is the one with the **highest variance**.



ISL Figure 6.15

How do we say this in math?

Let $\mathbf{X}$ be a data matrix with n samples, and p variables. From each variable, we subtract the mean of the column; i.e. we **center** the variables.

To find the first principal component $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})$, we solve the following optimization

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

Projection of the ith sample onto ϕ_1. Also known as **the score** z_{i1}

How do we say this in math?

Let $\mathbf{X}$ be a data matrix with n samples, and p variables. From each variable, we subtract the mean of the column; i.e. we **center** the variables.

p dims

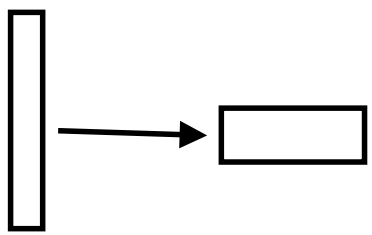
To find the first principal component $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})^t$, we solve the following optimization

Xi = (Xi1, ..., Xip)

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

ϕ1 is the first col of the matrix V't ϕ1 = V



variance
= ϕ1"t X"tX ϕ1
= ϕ1"tVD^2V"t ϕ1
a = V"t ϕ1
matrix a"t D^2a
= sum aj^2dj^2
aj is the largetst one,
dj is the second largest

sum aj = 1
a < di^2

Variance of the n samples projected onto ϕ_1.

How do we say this in math?

To find the second principal component $\phi_2 = (\phi_{12}, \ldots, \phi_{p2})$, we solve the following optimization

$$\max_{\phi_{12},\ldots,\phi_{p2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j2} x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} \phi_{j2}^2 = 1 \quad \text{and} \quad \sum_{j=1}^{p} \phi_{j1}\phi_{j2} = 0.$$

First and second principal components must be orthogonal.

along all the directions, we want to find the one maxmize the projections

How do we say this in math?

To find the second principal component $\phi_2 = (\phi_{12}, \ldots, \phi_{p2})$, we solve the following optimization

$$\max_{\phi_{12},\ldots,\phi_{p2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j2} x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} \phi_{j2}^2 = 1 \quad \text{and} \quad \sum_{j=1}^{p} \phi_{j1}\phi_{j2} = 0.$$

First and second principal components must be orthogonal.

Equivalent to saying that the scores $(z_{11}, \ldots, z_{n1})$ and $(z_{12}, \ldots, z_{n2})$ are uncorrelated.

Solving the optimization

This optimization is fundamental in linear algebra. It is satisfied by either:

- The singular value decomposition (SVD) of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Phi}^T$$

  where the ith column of $\mathbf{\Phi}$ is the ith principal component ϕ_i, and the ith column of $\mathbf{U}\mathbf{\Sigma}$ is the ith vector of scores $(z_{1i}, \ldots, z_{ni})$.

- The eigendecomposition of $\mathbf{X}^T\mathbf{X}$:

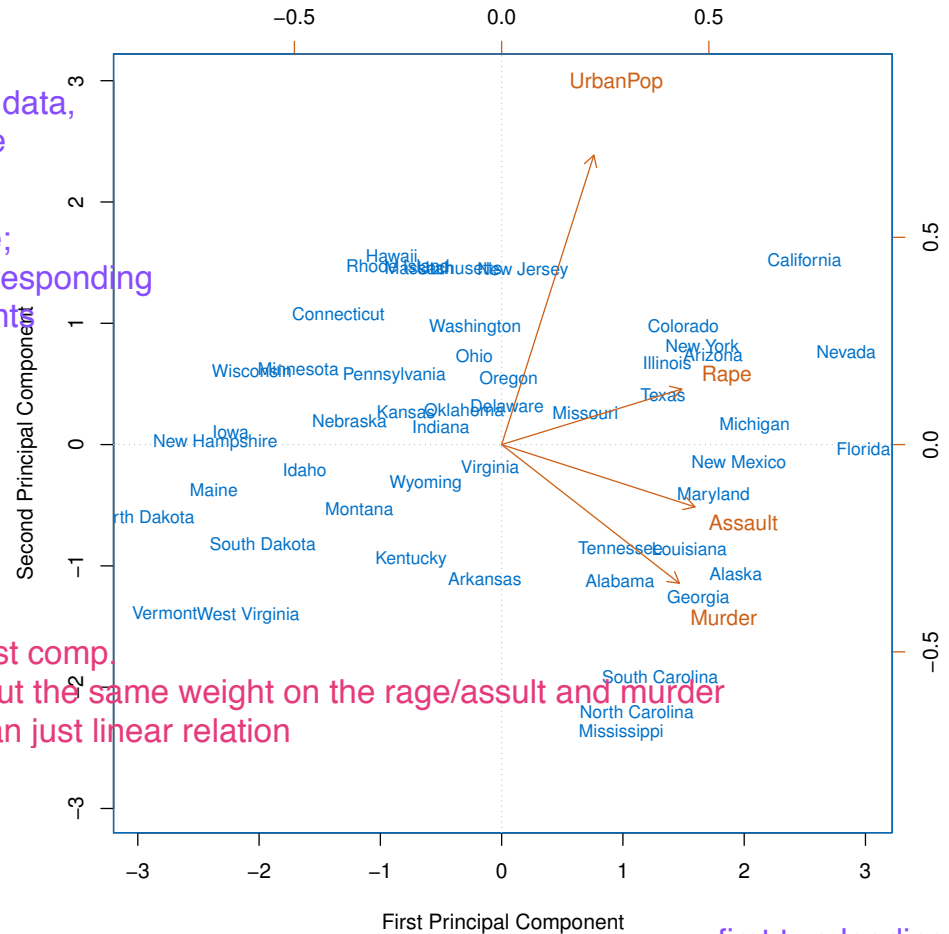
$$\mathbf{X}^T\mathbf{X} = \mathbf{\Phi}\mathbf{\Sigma}^2\mathbf{\Phi}^T$$

PCA in practice: The biplot

the plot of the first two variable projection; first two principles of components

blue points: projection of the data, each point represents a state

the vector means how much it weights on the first principle; vector of loading, it is the corresponding to the first principle component



it capture the variance of the first comp.
the first principle components put the same weight on the rage/assult and murder
they have close corre rather than just linear relation

first two loading of the first two components

ISL Figure 10.1

Scaling the variables

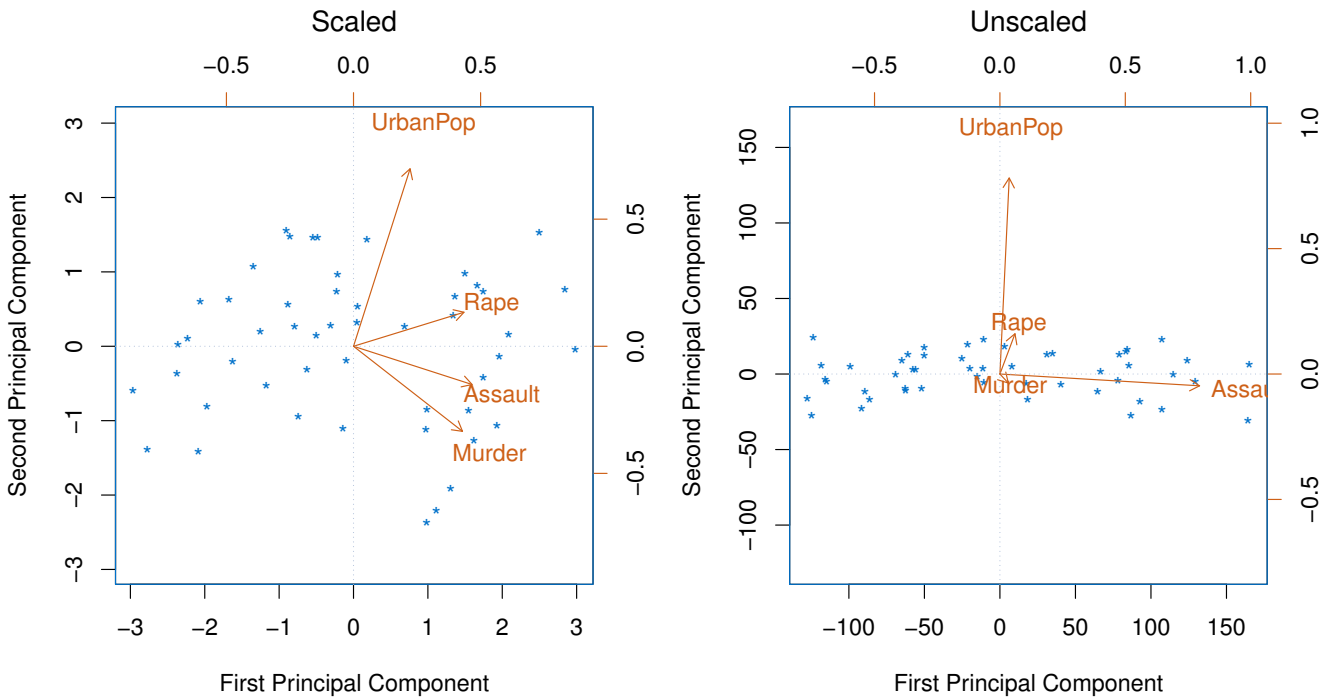
pca always center the data
and scale the variables

Most of the time, we don't care about the absolute numerical value of a variable. We care about the value relative to the spread observed in the sample.

Before PCA, in addition to **centering** each variable, we also multiply it times a constant to make its variance equal to 1.

if data is normalized, no need to scale
we always looke at the first two col(components)

Example: scaled vs. unscaled PCA



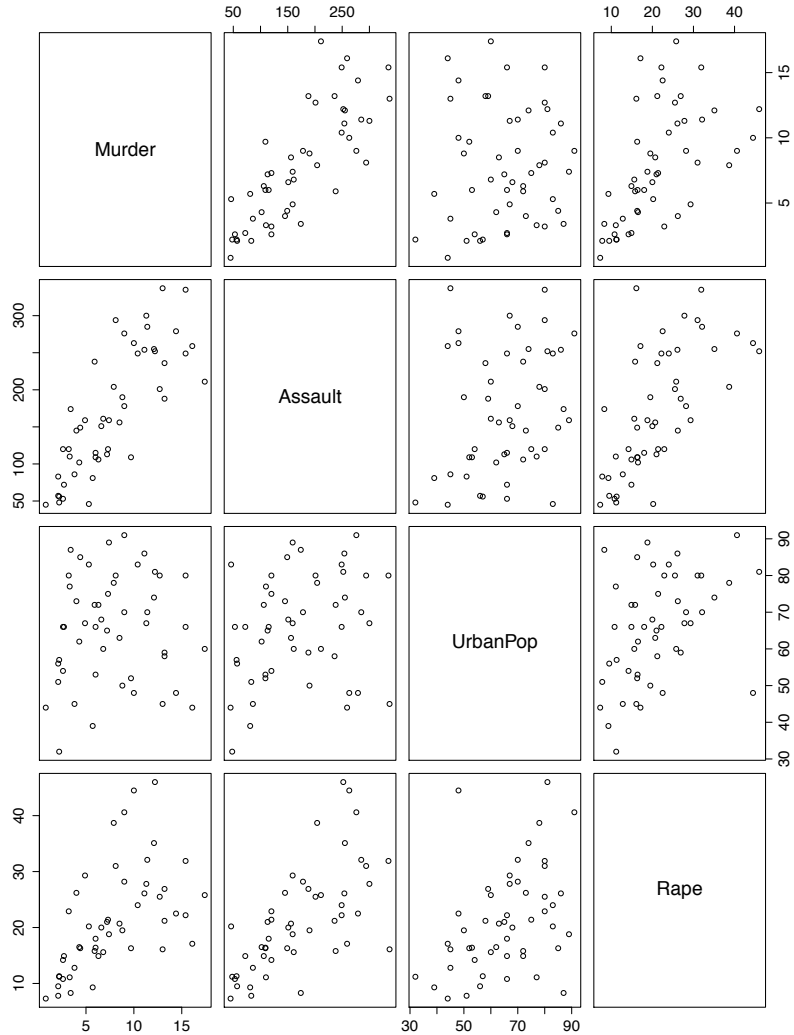
ISL Figure 10.3

Scaling the variables

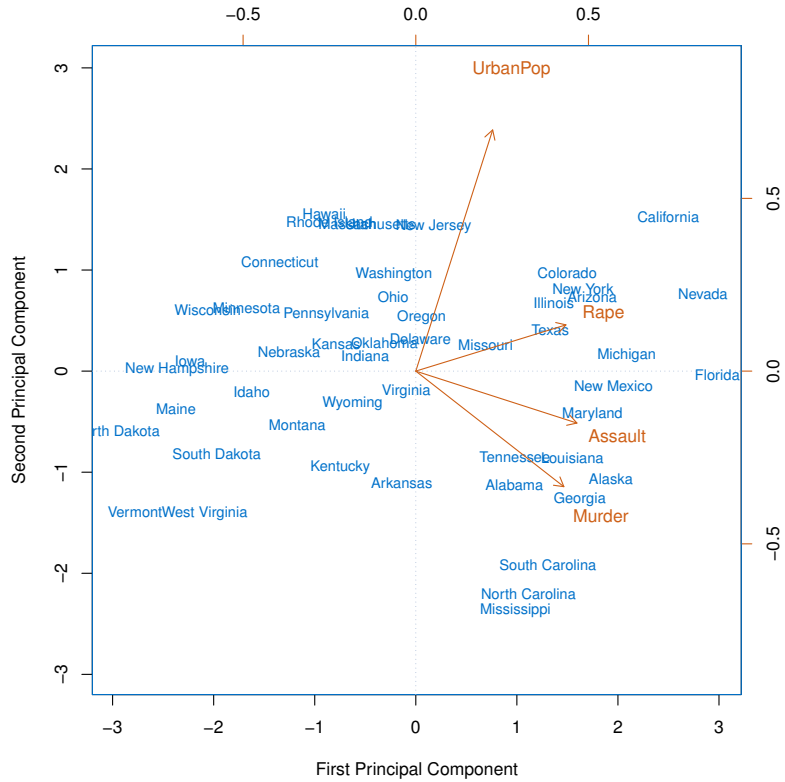
In special cases, we have variables measured in the same unit; e.g. gene expression levels for different genes.

Therefore, we care about the absolute value of the variables and we can perform PCA without scaling.

How many principal components are enough?



How many principal components are enough?



We said 2 principal components capture most of the relevant information. But how can we tell?

The proportion of variance explained

We can think of the top **principal components** as directions in space in which the data vary the most.

The ith **score vector** $(z_{1i}, \ldots, z_{ni})$ can be interpreted as a *new* variable. The variance of this variable decreases as we take i from 1 to p. However, the total variance of the score vectors is the same as the total variance of the original variables:

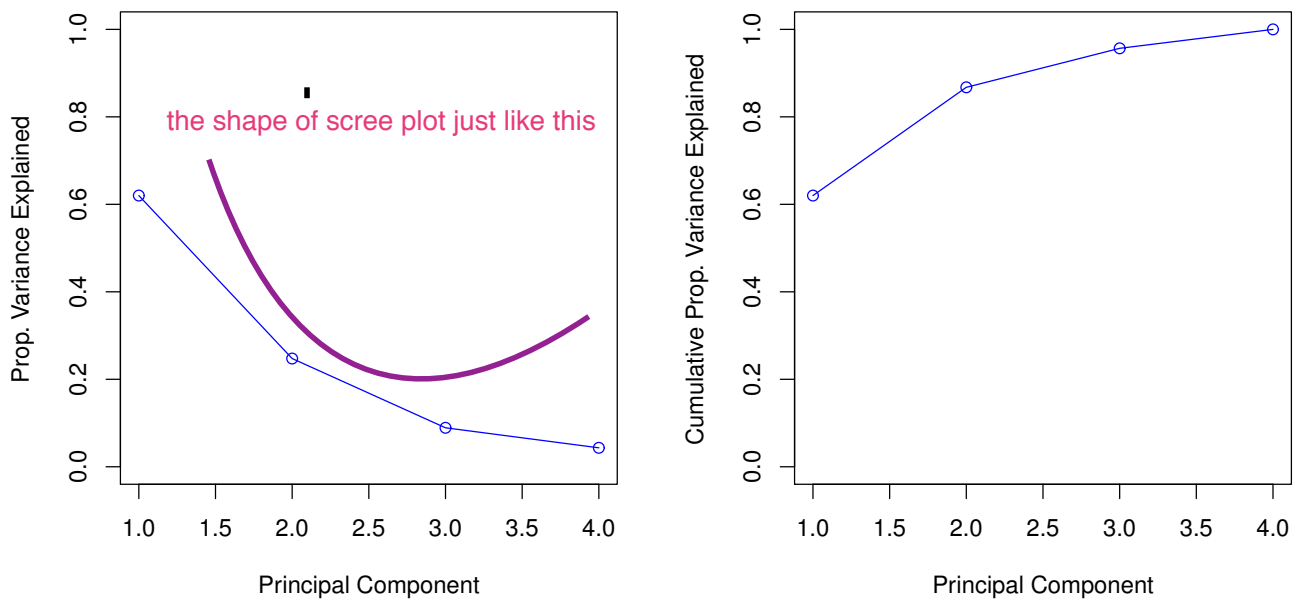
$$\sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} z_{ji}^2 = \sum_{k=1}^{p} \mathsf{Var}(x_k).$$

We can quantify how much of the variance is captured by the first m principal components/score variables.

The proportion of variance explained

The variance of the mth score variable is:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2 = \frac{1}{n}\mathbf{\Sigma}_{mm}^2.$$



Scree plot

Generalizations of PCA

PCA works under a Euclidean geometry in the space of variables. Often, the natural geometry is different:

- We expect some variables to be "closer" to each other that to other variables.
- Some correlations between variables would be more surprising than others.

Examples:

- Variables are pixel values, samples are different images of the brain. We expect neighboring pixels to have stronger correlations.
- Variables are rainfall measurements at different regions. We expect neighboring regions to have higher correlations.

Generalizations of PCA

There are ways to include this knowledge in a PCA. See:

1. Susan Holmes. *Multivariate Analysis, the French way.* (2006).
2. Omar de la Cruz and Susan Holmes. *An introduction to the duality diagram*. (2011).
3. Stéphane Dray and Thibaut Jombart. *Revisiting Guerry's data: Introducing spatial constraints in multivariate analysis.* (2011).
4. Genevera Allen, Logan Grosenick, and Jonathan Taylor. *A Generalized Least Squares Matrix Decomposition*. (2011).

Thanks to Sergio Bacallado and Peter Orbanz
for sharing the slides.