PRINCIPAL COMPONENT ANALYSIS

Dimensionality Reduction

In dimensionality reduction problems we observe

$$x_1, \ldots, x_n \in \mathbb{R}^d$$

The goal is to transform these inputs to new variables

$$x_i \longmapsto \theta_i \in \mathbb{R}^k$$

where $k < d$, in such a way that information loss is minimized.

Methods for dimensionality reduction may be categorized according to the following issues:

1) How is "information loss" quantified?

2) Supervised or unsupervised? If labels $y_1, \ldots, y_n$ are available, are they used?

3) Is the map $x \mapsto \theta$ linear or nonlinear?

4) Feature selection

(A) $\theta = \begin{bmatrix} & \\ & \end{bmatrix}$

versus feature extraction

$\theta = \begin{bmatrix} \\ \\ \end{bmatrix}$

PCA is described by

1)

2)

3)

4)

Linear Spans and Projections

Let $a_1, \ldots, a_k \in \mathbb{R}^d$ be linearly independent column vectors.

Definition The *linear span* of $a_1, \ldots, a_k$ is the set

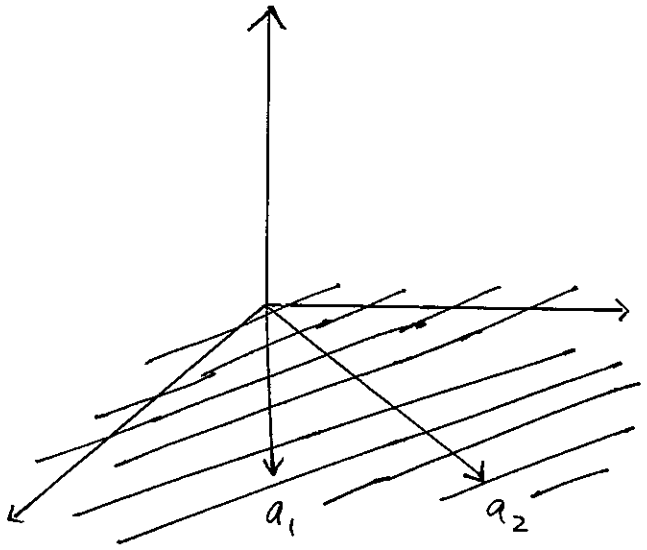
$$\left\{ x = \sum_{j=1}^{k} \theta^{(j)} a_j \; : \; \theta = [\theta^{(1)} \; \cdots \; \theta^{(k)}]^T \in \mathbb{R}^k \right\}$$

Equivalently, if we set

$$A = \begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix} \in \mathbb{R}^{d \times k}$$

then the linear span is the *image* or *column span* of A, and will be denoted $\langle A \rangle$.

$d = 3$
$k = 2$

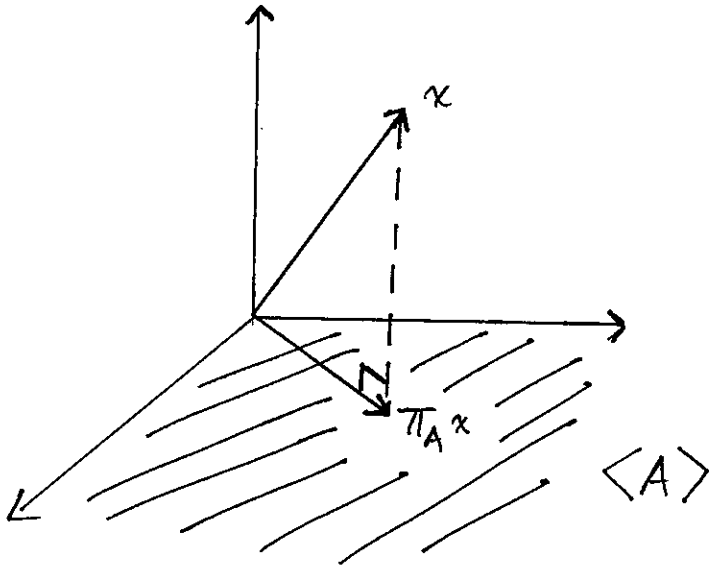


$\langle A \rangle$

Definition The mapping $\mathbb{R}^d \longrightarrow \langle A \rangle$ given by

$$\pi_A : x \longmapsto \text{closest point to } x \text{ in } \langle A \rangle$$

is called the <u>projection</u> onto $\langle A \rangle$.



Every point in $\langle A \rangle$ is equal to $A\theta$ for some $\theta \in \mathbb{R}^k$. Therefore $\pi_A x = A\hat{\theta}$ where

(B) $\hat{\theta} =$

Therefore

$$\Pi_A x = A (A^T A)^{-1} A^T x$$

If $a_1, \ldots, a_k$ are orthonormal, then

(c) $\Pi_A x =$

By the <u>orthogonality principal</u>, $x - \Pi_A x$ is orthogonal to every vector in $\langle A \rangle$.

<u>Exercise</u> | Verify this result.

Solution | Let $A\cdot\theta$ be an arbitrary vector in $\langle A \rangle$. We wish to show $\langle A\theta,\ x - \pi_A x \rangle = 0$. Now

$$\langle A\theta,\ x - \pi_A x \rangle = (A\theta)^T \left(x - A(A^TA)^{-1}A^T x \right)$$

(D) $=$

$=$

$=$

Projection matrices are also <u>idempotent</u>, meaning

$$\pi_A^{\,2} = \pi_A \cdot \pi_A$$

(E) $=$

Interpretation :

Principal Component Analysis

The idea behind PCA is to approximate

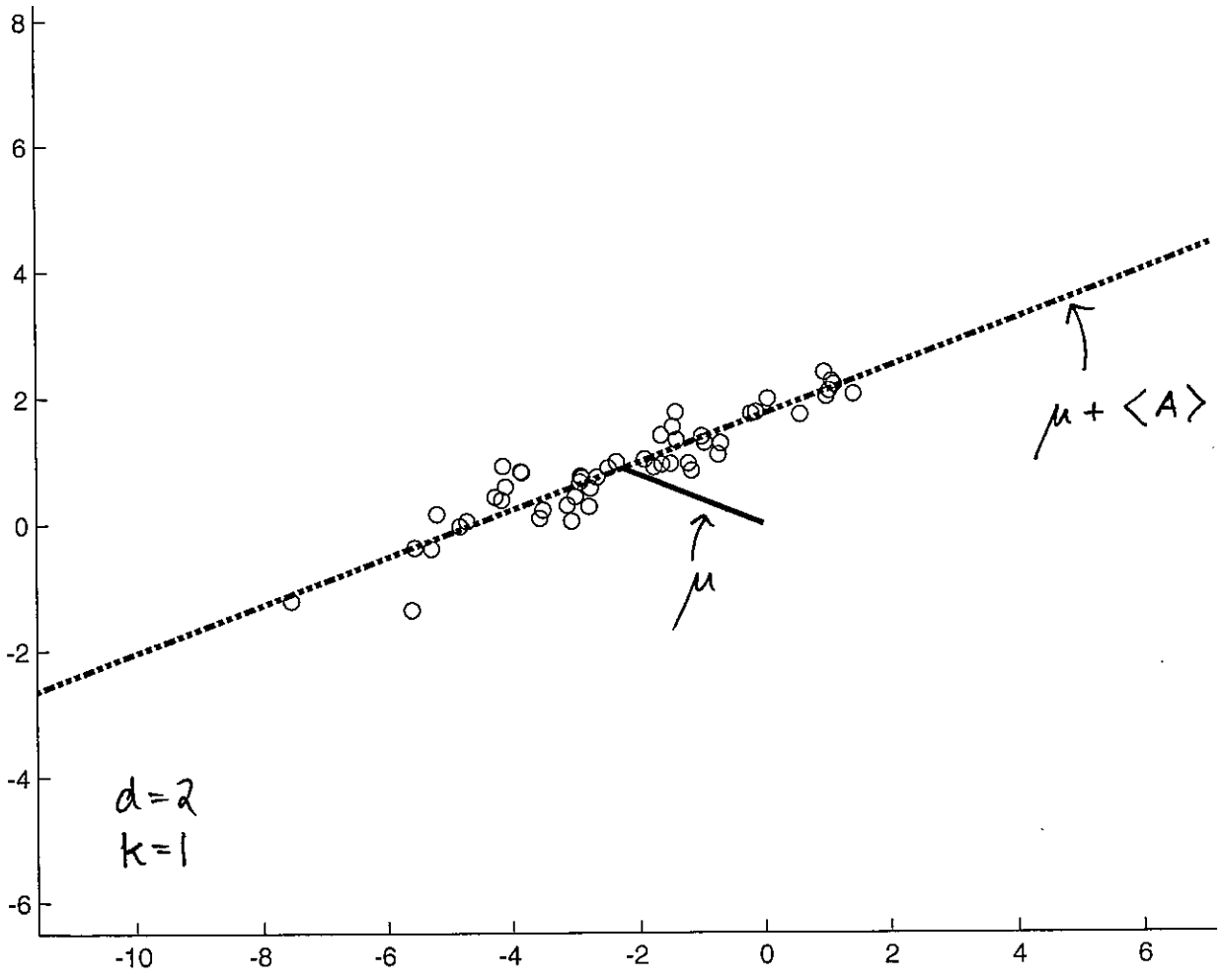
$$x_i \approx \mu + A\,\theta_i$$

where

$$\mu \in \mathbb{R}^d$$

$$A \in \mathbb{R}^{d \times k}, \text{ orthonormal columns}$$

$$\theta_i \in \mathbb{R}^k$$



Mathematically, we define μ, A, and $\theta_1, \ldots, \theta_n$ to be the solution of

$$\min_{\mu, A, \{\theta_i\}} \sum_{i=1}^{n} \| x_i - \mu - A\theta_i \|^2$$

The solution is given in terms of the <u>spectral representation</u> of

(F) $S =$

In particular, write

$$S = U \Lambda U^T$$

where

$$U = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{bmatrix}$$

with $U^T U = U U^T = I_{d \times d}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_d \geq 0$

Note: $S u_j = \lambda_j u_j \quad \forall j$

We will show that the optimal rank k linear approximation of the data is given by

(G) $\mu =$

$A =$

$\theta_i =$

Is A unique?

Terminology

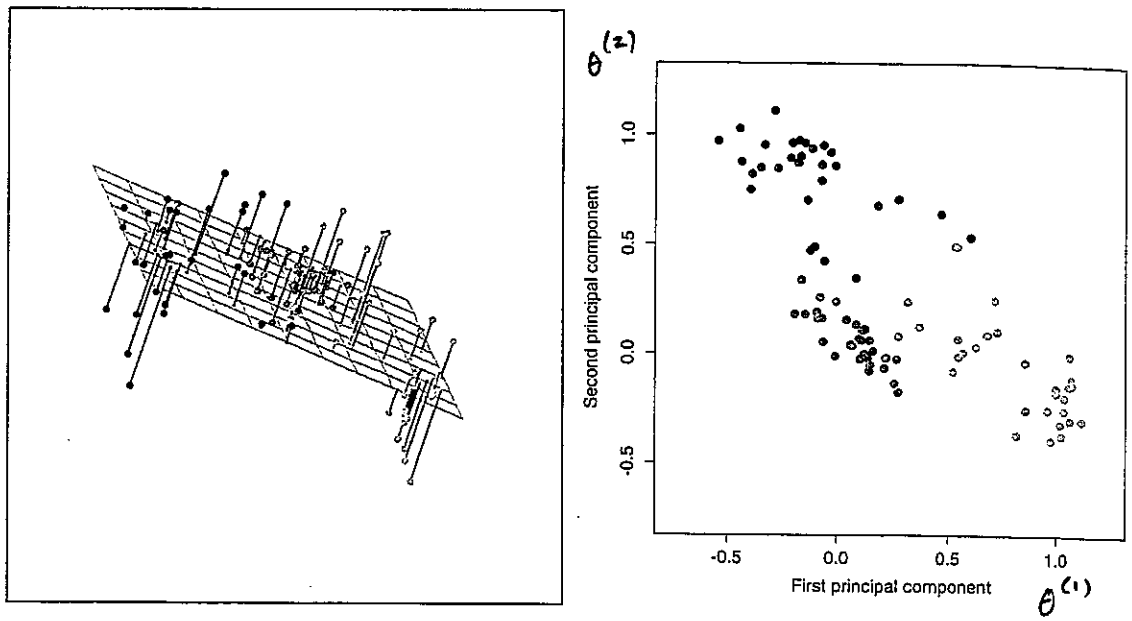
- principal component transformation

$$x \longmapsto \theta =$$

- jth principal component

$$\theta^{(j)} =$$

- jth principal eigenvector $\longrightarrow$



↑ Hastie, Tibshirani, & Friedman

FIGURE 14.21. *The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by [redacted], the first two principal components of the data.*

Proof: We wish to minimize

$$\sum_{i=1}^{n} \| x_i - \mu - A\theta_i \|^2$$

w.r.t $\mu, A, \{\theta_i\}$

Step 1: Eliminate $\{\theta_i\}$

Suppose A, μ are fixed. Then to minimize the expression we take

(H) $\theta_i =$

$=$

Step 2: Eliminate μ

Holding A fixed, we wish to minimize

$$\sum_{i=1}^{n} \| x_i - \mu - AA^T(x_i - \mu) \|^2$$

$$= \sum_{i} \| (I - AA^T)(x_i - \mu) \|^2$$

$$= \sum (x_i - \mu)^T \underbrace{(I - AA^T)^T (I - AA^T)}_{B} (x_i - \mu)$$

$$\frac{\partial}{\partial \mu} = -2 \sum_i B (x_i - \mu)$$

$$= -2B \cdot \sum_{i=1}^{n} (x_i - \mu) = 0$$

This is satisfied when

(I) $\mu =$

Step 3 Optimize A

It remains to minimize

$$\sum_{i=1}^{n} \| x_i - \bar{x} - A A^T (x_i - \bar{x}) \|^2$$

w.r.t. A

For convenience, assume $\bar{x} = 0$. (otherwise we could substitute $\tilde{x}_i = x_i - \bar{x}$)

Observe

$$\sum_i \|x_i - AA^T x_i\|^2 = \sum_i (x_i - AA^T x_i)^T (x_i - AA^T x_i)$$

$$= \sum_i x_i^T x_i - 2 x_i^T AA^T x_i + x_i^T AA^T AA^T x_i$$

(J) $=$

Therefore we can focus on <u>maximizing</u>

Introduce

$$\left.\begin{aligned} b_j &:= U^T a_j \\ y_i &:= U^T x_i \end{aligned}\right\} \text{ change of basis}$$

Our goal is to show

(K) $\langle A \rangle = \langle u_1, \ldots, u_k \rangle \iff b_j =$ [ ]

Now

$$\sum_i x_i^T A_k A_k^T x_i = \sum_i \| A^T x_i \|^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \left(a_j^T x_i \right)^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \left(a_j^T U U^T x_i \right)^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \left(b_j^T y_i \right)^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \left(b_j^T y_i \right) \left(y_i^T b_j \right)$$

$$= \sum_{j=1}^{k} b_j^T \left[\sum_{i=1}^{n} y_i y_i^T \right] b_j$$

Exercise | Simplify

Solution

$$\sum_{i=1}^{n} y_i y_i^T = \sum_{i=1}^{n} U^T x_i x_i^T U$$

$$= U^T \left[\sum_i x_i x_i^T \right] U$$

$$= U^T [nS] U$$

$$= n \cdot U^T \cdot U \Lambda U^T \cdot U$$

$$= n\Lambda$$ ∎

Thus, we need to maximize

$$\sum_{j=1}^{k} b_j^T \Lambda b_j$$

$$= \sum_{j=1}^{k} \sum_{l=1}^{d} \left(b_j^{(l)}\right)^2 \lambda_l$$

$$= \sum_{l=1}^{d} \left[\sum_{j=1}^{k} \left(b_j^{(l)}\right)^2 \right] \lambda_l$$

$$= \sum_{l=1}^{d} h_l \lambda_l \qquad \text{where } h_l = \sum_{j=1}^{k} \left(b_j^{(l)}\right)^2$$

Lemma: $0 \le h_l \le 1$ for $l = 1, \ldots, d$

and $\sum_{l=1}^{d} h_l = k$ (you'll show this on the homework)

Since $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$, we maximize

$$\sum_{\ell=1}^{d} h_\ell \lambda_\ell$$

by letting

$$h_\ell =$$

(L)

which implies

$$b_j^{(\ell)} = 0 \quad \text{for} \quad \ell > k$$

Therefore

$$a_j = U b_j \in \langle u_1, \ldots, u_k \rangle$$

▨

Maximum Variance Projections

PCA can be derived from a second perspective.

Suppose the data has zero mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.$$

What is the normal vector a_1 $(\|a_1\| = 1)$ for which

(M) $$\theta^{(1)} =$$

has maximal variance?

The variance of $\theta^{(1)}$ is defined to be

$$\mathrm{Var}(\theta^{(1)}) =$$

Theorem Assume $\lambda_1 > \lambda_2$.

If

$$\theta^{(1)} = a_1^T x$$

has maximal variance among all a_1 with $\|a_1\| = 1$, then

$$a_1 = 1\text{st principal eigenvector} = u_1$$

$$\theta^{(1)} = 1^{st} \text{ principal component}$$

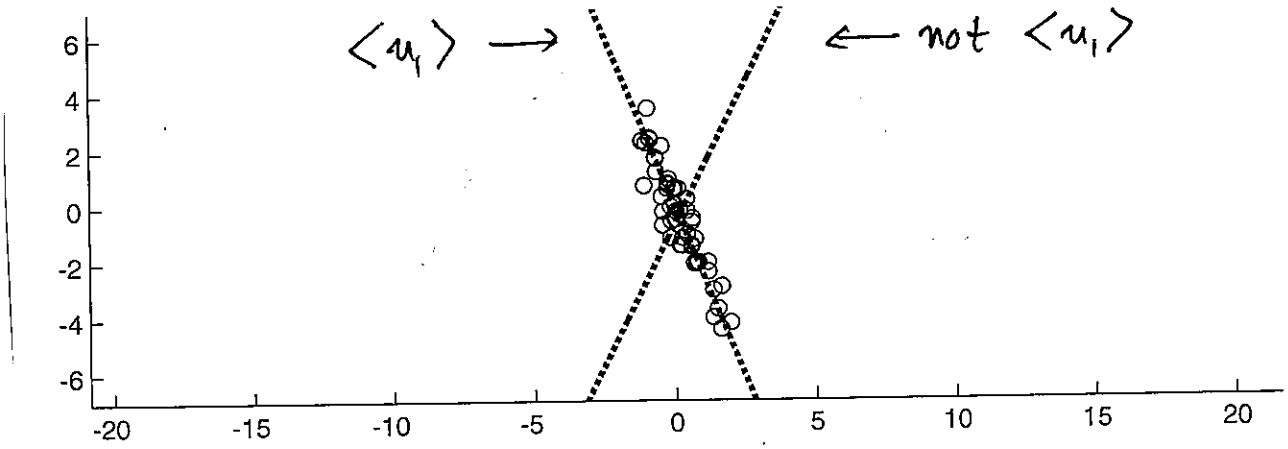
The variance of $u_1^T x$ is

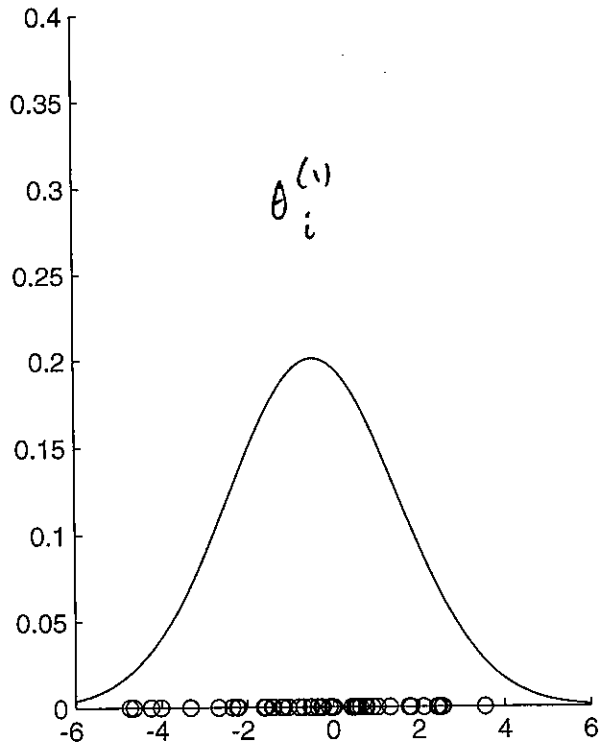
(N) $\mathrm{Var}(\theta^{(1)}) =$

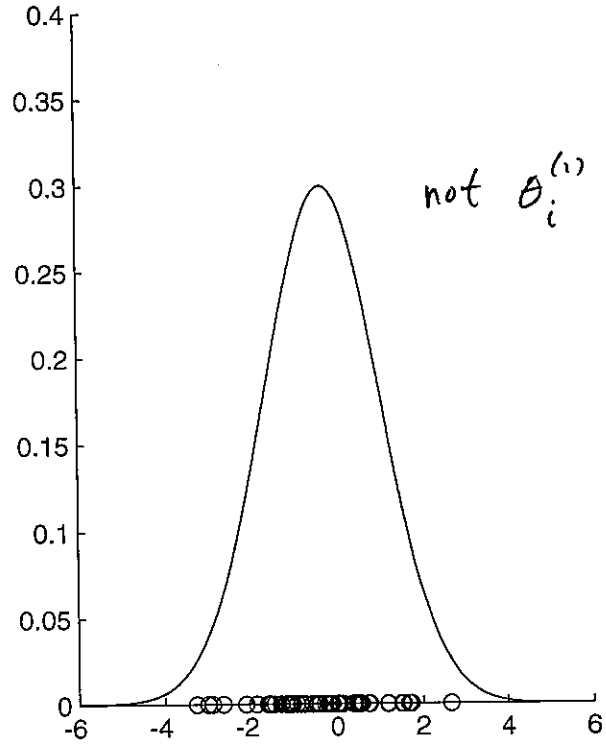
$=$

$=$

$=$







The other principal eigenvectors / components satisfy a similar property:

Theorem Assume $\lambda_k > \lambda_{k+1}$.

If

$$\theta^{(k)} = a_k^T x$$

has maximal variance among a_k such that

- $\| a_k \| = 1$
- $a_k \perp u_1, \ldots, u_{k-1}$

then

$a_k = k^{th}$ principal eigenvector $= u_k$

$\theta^{(k)} = k^{th}$ principal component

The variance of $\theta^{(k)}$ is _____.

Geometric Interpretation If the density of a random variable X is

$$\phi(x;\mu,\Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^d$$

then we say X is a multivariate Gaussian / normal random variable with parameters $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$, $\Sigma \succ 0$, and write $X \sim N(\mu, \Sigma)$.

If μ, Σ are unknown, we can estimate them from data $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \Sigma)$.

In particular, the Maximum Likelihood Estimate (MLE) of μ, Σ is obtained by maximizing the likelihood

$$l(\mu, \Sigma) := \prod_{i=1}^{n} \phi(x_i; \mu, \Sigma)$$

The maximum likelihood estimate of a Gaussian density is given by

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\Sigma} = S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

A contour of $\phi(x ; \bar{x}, S)$ consists of all points x such that

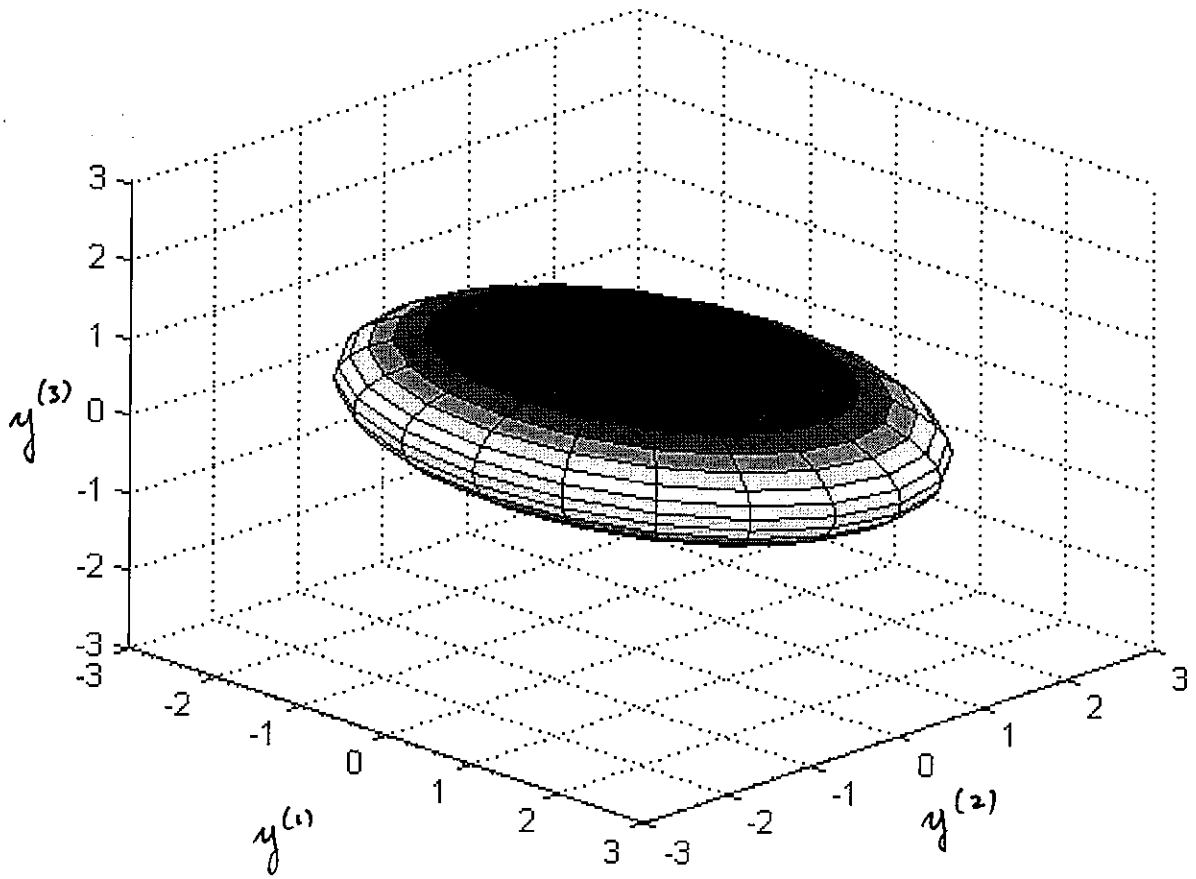
$$= c$$

Let $y = U^T x$, $\bar{y} = U^T \bar{x}$. We may write the contour-defining equation as

$$c =$$

In 3-d, the contour is an ellipsoid:

$$\frac{(y^{(1)} - \bar{y}^{(1)})^2}{\lambda_1} + \frac{(y^{(2)} - \bar{y}^{(2)})^2}{\lambda_2} + \frac{(y^{(3)} - \bar{y}^{(3)})^2}{\lambda_3} = c$$



$$\lambda_1 > \lambda_2 > \lambda_3$$

In general, PCA fits an elliptical distribution to the data, and projects onto the first k major axes of the resulting ellipsoid.

Practical Matters

Preprocessing

It is customary to center and scale a data set so that it has 0 mean and unit variance along each feature

For $j = 1$ to d

1) $\bar{x}^{(j)} = \frac{1}{n} \sum_{i=1}^{n} x_i^{(j)}$

$$x_i^{(j)} \leftarrow x_i^{(j)} - \bar{x}^{(j)}, \quad i = 1, \ldots, n$$

2) $\sigma^{(j)} = \left(\frac{1}{n} \sum_i \left(x_i^{(j)} \right)^2 \right)^{\frac{1}{2}}$

$$x_i^{(j)} \leftarrow x_i^{(j)} / \sigma^{(j)}, \quad i = 1, \ldots, n$$

End

This puts all features on an "equal playing field."

The steps may be omitted when

1) The data are known to be zero mean

2) The data are known to have comparable units of measurement.

Selecting k

It can be shown that

$$\min_{\mu, A_k, \{\theta_i\}} \sum_{i=1}^{n} \|x_i - \mu - A_k \theta_i\|^2 = n\left(\lambda_{k+1} + \cdots + \lambda_d\right)$$

When $k=0$ this becomes

$$\min_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|^2 = n\left(\lambda_1 + \cdots + \lambda_d\right),$$

which we call the total variation of the data.

Therefore, the % of total variation captured by k-dimensional PCA is

$$\%\,TV =$$

(P)

One heuristic for choosing k is to take the smallest k such that

$$\%\,TV \geq .95$$

↑

or .99, or some other number.

When to use PCA

- When the data form a single "point cloud" in space
- When the data are approximately Gaussian, or some other "elliptical" distribution.
- When low rank linear subspaces capture a majority of the variation

When not to use PCA

Key

A. Feature selection

$$\theta = \begin{bmatrix} x^{(2)} \\ x^{(5)} \\ x^{(17)} \\ \vdots \end{bmatrix}$$

Feature extraction

$$\theta = \begin{bmatrix} \varphi^{(1)}(x) \\ \varphi^{(2)}(x) \\ \vdots \end{bmatrix}$$

e.g., $\varphi^{(1)}(x) = (x^{(1)})^2 x^{(2)} + e^{-x^{(5)}}$

PCA: sum of squared errors, unsupervised, linear, feature extraction

B. $\hat{\theta} = \underset{\theta}{\arg\min} \frac{1}{2}\|x - A\theta\|^2 = (A^TA)^{-1}A^Tx$

C. $\Pi_A x = A \cdot A^T x$

D. $= \theta^T A^T (x - A(A^TA)^{-1}A^Tx)$

$= \theta^T (A^Tx - A^TA(A^TA)^{-1}A^Tx)$

$= \theta^T (A^Tx - A^Tx) = 0$

E. $\Pi_A^2 = \Pi_A \cdot \Pi_A = A(A^TA)^{-1}A^TA(A^TA)^{-1}A^T$

$= A(A^TA)^{-1}A^T = \Pi_A$

$\Rightarrow$ second projection has no effect

F. $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$, the sample covariance matrix

G. $\mu = \bar{x}$, $A = [u_1 \cdots u_k]$, $\theta_i = A^T(x_i - \bar{x})$

$x \mapsto \theta = A^T(x - \bar{x})$, $\theta^{(j)} = u_j^T(x - \bar{x})$, u_j

H. $(A^TA)^{-1}A^T(x_i-\mu) = A^T(x_i-\mu)$ I. $\mu = \bar{x}$

J. $\sum_i x_i^T x - x_i^T AA^T x_i$, $\sum_i x_i^T AA^T x_i$

K. $b_j = [\text{whatever} \mid \underbrace{0 \cdots 0}_{k+1 \leftrightarrow d}]^T$

L. $h_\ell = \begin{cases} 1 & \text{if } \ell \le k \\ 0 & \text{if } \ell > k \end{cases}$

M. $\theta^{(1)} = a_1^T x$, $Var(\theta^{(1)}) := \frac{1}{n}\sum_{i=1}^{n} (a_1^T x_i)^2$

N. $Var(\theta^{(1)}) = \frac{1}{n}\sum (u_1^T x_i)^2 = \frac{1}{n}\sum (u_1^T x_i)(x_i^T u_1)$

$= u_1^T \cdot \frac{1}{n}\sum x_i x_i^T \cdot u_1$

$= u_1^T \cdot U\Lambda U^T \cdot u_1$

$= [1\ 0 \cdots 0] \cdot \Lambda \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \lambda_1$

O. $c = (x-\bar{x})^T S^{-1}(x-\bar{x}) = (u_y - u_{\bar{y}})^T S^{-1}(u_y - u_{\bar{y}})$

$= (y-\bar{y})^T U^T \cdot U\Lambda^{-1}U^T \cdot U(y-\bar{y}) = (y-\bar{y})^T \Lambda^{-1}(y-\bar{y})$

$= \sum_{j=1}^{d} \frac{(y^{(j)} - \bar{y}^{(j)})^2}{\lambda_j}$

P. $\%TV = \dfrac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_d}$