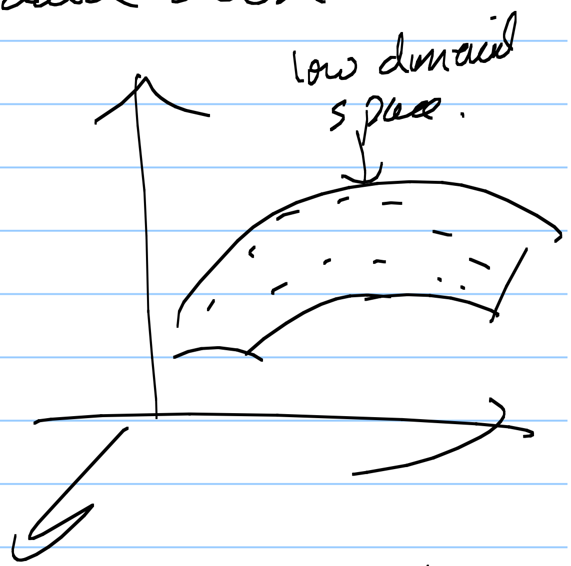
Note Title 10/15/2006

Principal Component Analysis (PCA)

One way to deal with the curse of dimensionality is to project data down onto a space of low dimensions.

There are a number of different techniques for doing this → eg. multidimensional scaling. Too many to deal with in this course.



Now we discuss the most basic method — Principal Component Analysis. (PCA)

CONVENTION : $\underline{\mu}^T\underline{\mu}$ is a scalar $\mu_1^2+\mu_2^2+\dots+\mu_D^2$

$\underline{\mu}\,\underline{\mu}^T$ is a matrix $\begin{pmatrix} \mu_1^2 & \mu_1\mu_2 & \mu_1\mu_3 & \dots \\ & \mu_2^2 & & \\ & & \ddots & \end{pmatrix}$

(2) (N.B. different convention than blackboard notes – but same as in book)

Data samples $\underline{x}_1, \dots, \underline{x}_N$ in D-dim space.

Compute the mean $\mu = \frac{1}{N}\sum_{i=1}^{N} \underline{x}_i$.

Compute the covariance:

$$\underline{\underline{K}} = \frac{1}{N}\sum_{i=1}^{N} (\underline{x}_i-\underline{\mu})(\underline{x}_i-\underline{\mu})^T.$$

Next compute the eigenvalues and eigenvectors of $\underline{\underline{K}}$

Solve $\underline{\underline{K}}\,\underline{e} = \lambda\,\underline{e}$

$\lambda_1 > \lambda_2 > \dots > \lambda_N$

note: $\underline{\underline{K}}$ is symmetric – so eigenvalues are real, eigenvectors are orthogonal.

PCA reduces the dimension by by keeping the eigenvectors $\underline{e}_i$ with $\lambda_i < T$

threshold

Let M eigenvectors be kept.

Then project data $\underline{x}$ onto the subspace spanned by the first M eigenvectors. (After substracting out the mean).

(3.) **Formally:**

Project.

$$\underline{x} - \underline{\mu} = \sum_{\nu=1}^{D} a_\nu \underline{e}_\nu$$

where the coefficients are given by

$$a_\nu = (\underline{x} - \underline{\mu}) \cdot \underline{e}_\nu$$

(orthogonality, means $\underline{e}_\nu \cdot \underline{e}_\mu = \delta_{\nu\mu}$ — Kronecker delta

Here

$$\underline{x} = \underline{\mu} + \sum_{\nu=1}^{D} \left\{ (\underline{x} - \underline{\mu}) \cdot \underline{e}_\nu \right\} \underline{e}_\nu$$

no dimension reduction
(no compression)

Then, approximate

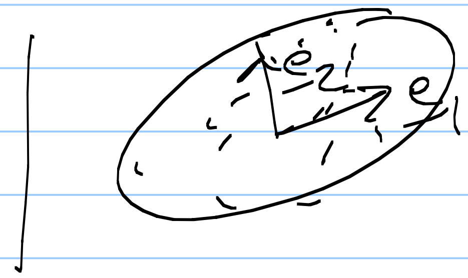
$$\underline{x} \approx \underline{\mu} + \sum_{\nu=1}^{M} \left\{ (\underline{x} - \underline{\mu}) \cdot \underline{e}_\nu \right\} \underline{e}_\nu.$$

Projects the data onto the M-dim subspace.

$$\underline{\mu} + \sum_{\nu=1}^{M} b_\nu \underline{e}_\nu \quad .//$$

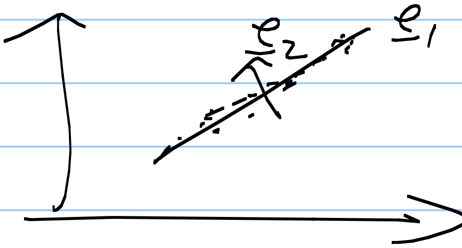
(4) In 2-dimensions

Visually



The eigenvectors of $\underline{\underline{K}}$ correspond to the second order moments of the data.

If the data lies (almost) on a straight line, then $\lambda_1 \gg 0$, $\lambda_2 \simeq 0$



PCA and Gaussian Distribution

PCA is equivalent to performing ML estimation of the parameters of a Gaussian

$$p(\underline{x}|\underline{\mu},\underline{\underline{\Sigma}}) = \frac{1}{\sqrt{2\pi}\sqrt{\det \underline{\underline{\Sigma}}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu})}.$$

to get $\hat{\underline{\mu}}, \hat{\underline{\underline{\Sigma}}}$. And then throw away the directions where the variance is small.

(5) <u>Cost Function for PCA.</u>

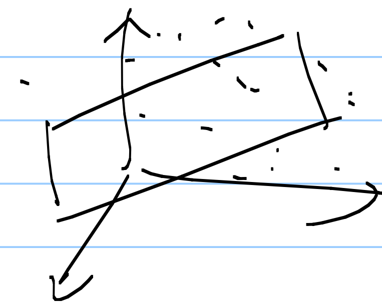
$$J(\underline{\mu}, \{a\}, \{e\}) = \sum_{k=1}^{N} \| (\underline{\mu} + \sum_{i=1}^{M} a_{ki} \underline{e}_i) - \underline{x}_k \|^2$$

<u>Minimize</u> J w.r.t. $\underline{\mu}, \{a\}, \{e\}$.

Data $\{ \underline{x}_k : k = 1 \text{ to } N \}$.

The $\{a_{ki}\}$ are projection coefficients.

<u>Intuition</u>: find the M-dimensional subspace s.t. the projections of the data onto this subspace have minimal error.



<u>Minimizing</u> J, gives the $\{\hat{\underline{e}}_i\}$'s to be the eigenvectors of the covariance matrix $\underline{\underline{K}} = \frac{1}{N} \sum_{k=1}^{N} (\underline{x}_k - \underline{\mu})(\underline{x}_k - \underline{\mu})^T$

$$\underline{\mu} = \frac{1}{N} \sum_{k=1}^{N} \underline{x}_k$$

$$\hat{a}_{ki} = (\underline{x}_k - \hat{\underline{\mu}}) \cdot \hat{\underline{e}}_i$$ the projection coefficients.

(6) To understand this fully, you must understand Singular Value Decomposition (SVD)

We can re-express the criteria as

$$J[\underline{\mu}, \{a\}, \{e\}] = \sum_{k=1}^{N} \sum_{b=1}^{D} \left\{ (\mu_b - x_{bk}) + \sum_{i=1}^{M} a_{ki} e_{ib} \right\}^2$$

where b denotes the vector components.

This is an example of a general class of problem.

Let $$E[\psi, e] = \sum_{a=1, k=1}^{a=D, k=N} \left(\tilde{x}_{ak} - \sum_{\nu=1}^{M} \psi_{a\nu} \phi_{\nu k} \right)^2$$

Goal: minimize $E[\psi, e]$ w.r.t. ψ, e.

This is a bilinear problem, that can be solved by SVD.

Note: $\tilde{x}_{ak} = x_{ak} - \mu_a$

the position of the point, relative to the mean.

Note: $\underline{\underline{X}}$ is not a square matrix. So it has no eigenvalues or eigenvectors.

(7) SVD

We can express any $N \times D$ matrix $\underline{\underline{X}}$ X_{ak} in form

$$\underline{\underline{X}} = \underline{\underline{E}}\ \underline{\underline{D}}\ \underline{\underline{F}}$$

$$X_{ak} = \sum_{\mu,\nu=1}^{M} e_{a\mu}\, d_{\mu\nu}\, f_{\nu k}$$

where $\underline{\underline{D}} = \{d_{\mu\nu}\}$ is a diagonal matrix ($d_{\mu\nu}=0, \mu\neq\nu$)

$$\underline{\underline{D}} = \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_M} \end{pmatrix}$$

, where the $\{\lambda_i\}$ are eigenvalues of $\underline{\underline{X}}\,\underline{\underline{X}}^T$ (equivalently of $\underline{\underline{X}}^T\underline{\underline{X}}$).

μ & ν label the eigenvectors.

$\underline{\underline{E}} = \{e_{a\mu}\}$ are eigenvectors of $(\underline{\underline{X}}\,\underline{\underline{X}}^T)_{ab}$

$\underline{\underline{F}} = \{f_{\nu k}\}$ are eigenvectors of $(\underline{\underline{X}}^T\underline{\underline{X}})_{kl}$

Note: For $\tilde{\underline{\underline{X}}}$ defined on previous page, we get that $(\tilde{\underline{\underline{X}}}\,\tilde{\underline{\underline{X}}}^T) = \sum_{k=1}^{N} (\underline{x}_k - \underline{\mu})(\underline{x}_k - \underline{\mu})^T$

Note: If $(\underline{\underline{X}}\,\underline{\underline{X}}^T)\underline{e} = \lambda\, \underline{e}$

then $(\underline{\underline{X}}^T\underline{\underline{X}})(\underline{\underline{X}}^T\underline{e}) = \lambda(\underline{\underline{X}}^T\underline{e})$

This relates the eigenvectors of $\underline{\underline{X}}\,\underline{\underline{X}}^T$ and of $\underline{\underline{X}}\,\underline{\underline{X}}^T$.
(Calculate the eigenvectors for the smallest matrix, then deduce those of the bigger matrix - $D < N$)

(8)

Minimize:

$$E[\Psi, e] = \sum_{a=1,k=1}^{a=D,k=N} \left(\tilde{X}_{ak} - \sum_{\nu=1}^{M} \Psi_{a\nu}\, \phi_{\nu k}\right)^2$$

we set.

$$\begin{cases} \Psi_{a\nu} = \sqrt{d_{\nu\nu}}\; e_a^{\nu} \\ \phi_{\nu k} = \sqrt{d_{\nu\nu}}\; f_k^{\nu} \end{cases}$$

Take M biggest terms in the SVD expansion of $\underline{\underline{X}}$.

<u>But</u> there is an ambiguity.

$$\sum_{\nu=1}^{M} \Psi_{a\nu}\, \phi_{\nu k} = \underline{\underline{\Psi}}\; \underline{\underline{\phi}}_{ak}$$

matrix multiplied.

$$= \underline{\underline{\Psi}}\; \underline{\underline{A}}\; \underline{\underline{A}}^{-1} \underline{\underline{\phi}}_{ak}$$

for any $M \times M$ invertible matrix $\underline{\underline{A}}$.

$$\underline{\underline{\Psi}} \to \underline{\underline{\Psi}}\,\underline{\underline{A}}$$

$$\underline{\underline{\phi}} \to \underline{\underline{A}}^{-1}\underline{\underline{\phi}}$$

This gets rid of the ambiguity.

For the PCA problem – we have constraints that the projection directions are orthogonal unit eigenvectors

(9) <u>Relate</u> SVD to <u>PCA</u> <u>(Linear Albgebra)</u>

Start with an $n \times m$ matrix $\underline{\underline{X}}$.

$\underline{\underline{X}}\,\underline{\underline{X}}^T$ is a symmetric $n \times n$ matrix

$\underline{\underline{X}}^T\underline{\underline{X}}$ is a symmetric $m \times m$ matrix.

$\left((\underline{\underline{X}}\,\underline{\underline{X}}^T)^T = \underline{\underline{X}}\,\underline{\underline{X}}^T \right)$

By standard linear algebra.

$\underline{\underline{X}}\,\underline{\underline{X}}^T \underline{e}^\mu = \lambda^\mu \underline{e}^\mu$ $\quad$ n eigenvalues λ^μ
eigenvectors $\underline{e}^\mu$

eigenvectors are orthogonal $\underline{e}^\mu \cdot \underline{e}^\nu = \delta^{\mu\nu}$.

Similarly $\underline{\underline{X}}\,\underline{\underline{X}}^T \underline{f}^\nu = \tau^\nu \underline{f}^\nu$ $\quad$ m eigenvalues τ^ν
eigenvectors $\underline{f}^\nu$
$\underline{f}^\mu \cdot \underline{f}^\nu = \delta^{\mu\nu}$.

The $\{\underline{e}^\mu\}$ and $\{\underline{f}^\mu\}$ are related because

$$(\underline{\underline{X}}^T\underline{\underline{X}})(\underline{\underline{X}}^T\underline{e}^\mu) = \lambda^\mu(\underline{\underline{X}}^T\underline{e}^\mu)$$

$$(\underline{\underline{X}}\,\underline{\underline{X}}^T)(\underline{\underline{X}}\,\underline{f}^\mu) = \tau^\mu(\underline{\underline{X}}\,\underline{f}^\mu)$$

<u>Hence:</u> $\underline{\underline{X}}^T\underline{e}^\mu \propto \underline{f}^\mu$, $\quad \underline{\underline{X}}\,\underline{f}^\mu \propto \underline{e}^\mu$ $\quad \lambda^\mu = \tau^\mu$

If $n > m$, then there are n eigenvectors $\{\underline{e}_\mu\}$ and m eigenvectors $\{\underline{f}_\mu\}$. So some $\underline{f}_\mu$ relate to several $\{\underline{e}_\mu\}$.

(10). Claim: we can express

$$\mathbf{X} = \sum_\mu \alpha^\mu \mathbf{e}^\mu \mathbf{f}^{\mu T}$$

$$\mathbf{X}^T = \sum_\mu \alpha^\mu \mathbf{f}^\mu \mathbf{e}^{\mu T}$$

for some α^μ (we will solve for α^μ later.)

Verify the claim

$$\mathbf{X}\mathbf{f}^\nu = \sum_\mu \alpha^\mu \mathbf{e}^\mu \mathbf{f}^{\mu T} \mathbf{f}^\nu$$

$$= \sum_\mu \alpha^\mu \delta_{\mu\nu} \mathbf{e}^\mu = \alpha^\nu \mathbf{e}^\nu \,.//$$

$$\mathbf{X}\mathbf{X}^T = \sum_{\mu,\nu} \alpha^\nu \mathbf{e}^\nu \mathbf{f}^{\nu T} \alpha^\mu \mathbf{f}^\mu \mathbf{e}^{\mu T}$$

$$= \sum_{\mu,\nu} \alpha^\nu \alpha^\mu \mathbf{e}^\nu \delta_{\mu\nu} \mathbf{e}^{\mu T} = \sum_\mu (\alpha^\mu)^2 \mathbf{e}^\mu \mathbf{e}^{\mu T}.$$

Similarly $\mathbf{X}^T\mathbf{X} = \sum_\mu (\alpha^\mu)^2 \mathbf{f}^\mu \mathbf{f}^{\mu T}$, so $(\alpha^\mu)^2 = \lambda^\mu$ //

(Because we can express a symmetric matrix in form $\sum_\mu x_\mu \mathbf{e}^\mu \mathbf{e}^{\mu T}$ //

$$\mathbf{X} = \sum_\mu \alpha^\mu \mathbf{e}^\mu \mathbf{f}^{\mu T} \quad \text{is the SVD of } \mathbf{X}$$

In coordinates: $X_{ai} = \sum_\mu \alpha^\mu e^\mu_a f^\mu_i$

$$X_{ai} = \sum_{\mu,\nu} e^\mu_a \alpha^\mu \delta_{\mu\nu} f^\nu_i$$

$$\mathbf{X} = \mathbf{E}\,\mathbf{D}\,\mathbf{F} \qquad E_{a\mu} = e^\mu_a,\ D_{\mu\nu} = \alpha^\mu \delta_{\mu\nu}$$

$$F_{\nu i} = f^\nu_i$$

(11) <u>Effectiveness of PCA.</u>

In practice, PCA is often effective at (minimising) reduction of data dimension.

<u>But</u> it will not be effective for some problems.

For example, if the data is a set of strings

$$(1, 0, 0, 0 \dots) = \underline{x}_1$$
$$(0, 1, 0, 0 \dots) = \underline{x}_2$$

$$(0, 0, 0, 0 \dots, 0, 1) = \underline{x}_N$$

then the eigenvalues do not fall off as PCA requires.